A Matching Algorithm for Facial Memory Recall in Forensic Applications

by

LAU Kwok Kin

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy

in

Systems Engineering and Engineering Management

May 2000

To God, my parents and Hazel.

Abstract

In the literature of facial recognition, matching is performed on gray-level photographic faces. However, in facial recall, human's face is usually represented as an image of sketch or a composite of facial features. Its subsequent facial matching is now different from the traditional photo-to-photo matching and extends to sketch-to-sketch, sketch-to-photo and photo-to-sketch. In this thesis, we attempt to deal with these kinds of matching, based on a self-developed forensic application software of facial recall. With this software, a face is made up of the templates of sketches of different facial features. In the first part of the thesis, we propose a component-based encoding scheme to provide a simple and efficient way for the process of the sketch-to-sketch matching. This scheme pre-compares the facial features and stores the similarity values into a similarity matrix for each feature class. Matching is simply reduced to a summation process so that the required time becomes negligible. Also, we have done experiments to investigate the effect on matching when varying major facial features, such as eyes, eyebrows, nose... etc., individually; In the second part, we attempt to apply two common face recognition techniques, Principal Component Analysis (PCA) and Local Feature Analysis (LFA), in the problem of sketch-to-photo/photo-to-sketch matching. Experiments are done to compare the performance of the two algorithms.

摘要

在面部識別的研究中，比較所用的圖像都是灰度相片，但是，在面部重組的過程裡，人的容貌都是以素描圖像或拼圖來表達。這樣，隨後所進行的面部比較便和傳統的〝相片與相片〞的比較有所不同，並伸展至〝素描與素描〞、〝素描與相片〞和〝相片與素描〞的比較。在本論文中，我們嘗試以一個自行開發的面部重組應用軟件作基礎去處理這些新的問題。利用這軟件，人的容貌便由不同的面部特徵的素描模板所組成。在本文的前半部，我們建議了一個成分爲本的編碼方案，爲隨後的〝素描與素描〞比較提供了一個簡單快捷的方法。這方案預先比較了同類型的面部特徵，然後將它們之間的相似值儲存在一個相似度矩陣。這樣，比較便可簡化爲一個計算總和的程序。除此之外，我們進行了一些實驗去研究某些主要特徵（如眼睛、眼眉、鼻子等）的大小改變，如何影響面部比較的結果。在本文的後半部，我們嘗試將主要成份分析（PCA）和局部特徵分析（LFA）應用在〝素描與相片〞和〝相片與素描〞的比較上，並進行了一些實驗去比較它們在處理這些問題上的表現。

Acknowledgements

I would like to give thanks to all the people that have offered their help to me over the last two and a half years.

I particularly would like to thank my supervisor, Professor Lam Kai Pui, for his kind and patient guidance throughout my research. Without his invaluable advice and endless encouragement, this paper may never be completed.

I would like to thank my colleagues in the Department of SEEM. I am grateful to Chu Sir who teaches me a lot in using LaTeX and works with me everyday until nearly midnight; Terence Yung who always encourages me; Elvis Law who brings me a lot of fun and Murphy Yip works with me to develop AICAMS-FIT. I am also grateful to Philip Woo, Francis Lam, Joseph Lee, Matt Chan, Lam Ngok, Timothy Chan, Keung Chi Kin and Smooch Tang who give me a happy research life.

Lastly, I would like to specially thank the police artist, Wong Kwok Yao, who provides me the images of facial sketch and facial photo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past 20 years, human and machine recognition of faces has been an active research area conducted in engineering, neuroscience and psychophysics, spanning several disciplines such as image processing, pattern recognition, computer vision and neural network. Many face recognition technologies have been developed and applied to numerous commercial and law enforcement applications. These applications range from static matching of controlled format photographs such as passports, credit cards, photo ID's, driver's license, and mug shots to real-time matching of surveillance video images presenting different constraints in terms of processing requirements [1].

Most of the face recognition technologies, as well as applications, deal with human faces of gray-level frontal images. These images are usually captured from photographs. However, in some situations, a human face may be represented by a sketch or a line drawing instead. Moreover, it may be represented as a composite of facial features. Facial memory recall associated with eyewitness identification in forensic procedures is one of the examples.

In [2], forensic tasks can be roughly characterized as production (recall) or match-

ing (recognition). Recall-like activities aim at getting a description or representation of the facial information in an eyewitness' memory. For examples, generating verbal descriptions or hard-copy representations of faces. While recognition-like activities aim at finding a face from external sources, such as photos or composite faces, that matches the face in a witness's memory. For examples, searching a mugfile, reviewing a lineup, examining a photospread, and identifying (positively or negatively) an individual in a showup or in a courtroom.

Let's consider the following forensic task of facial recall: the generation of hard-copy representations of faces. In this task, the witness is asked to compose a picture of a culprit using a library of facial features such as noses, eyes, lips, hair, ears, etc. For each features, there may be different sizes and shapes such as large, small, pointed, curly, etc., from which one that is closest to witness's memory recall is chosen. Those facial features are usually in line drawing format (sketch) rather than photographic format. In this situation, the human face is represented as a composite of sketches of facial features. Now, the subsequent facial matching problem, such as the mugfile search in the forensic setting, is quite different from the traditional "photo-to-photo" matching. It becomes a "sketch-to-sketch" matching if the searching database contains facial sketches[1] or a "sketch-to-photo" matching if the searching database contains facial photos only, provided that the test face is a facial sketch. If the test face is a facial photo while the searching database contains facial sketches only, then the matching will become a "photo-to-sketch" matching.

In the literature, however, there is a lack of more in-depth study on such kinds of facial matching. The related study we can find is reported in [3]. It is a case study conducted by W. Konen. He compares facial line drawings with gray-level images

[1]Throughout the thesis, the term "facial sketch" is used to describe a facial image in sketch format, while the term "facial photo" is used to describe a facial image in photographic format.

using a software tool called PHANTOMAS. The algorithm used in PHANTOMAS is based on von der Malsburg's Elastic Graph Matching Algorithm. It is reported that the algorithm successfully recognize faces from facial line drawings.

1.1 Objective of This Thesis

In this research, we works on above kinds of matching based on a self-developed forensic application software of facial recall called AICAMS-FIT. For the sketch-to-sketch matching, we propose a component-based encoding scheme which provides an efficient matching. For the sketch-to-photo/photo-to-sketch matching, we attempt to use Principal Component Analysis (PCA) and Local Feature Analysis (LFA) and compare their performance.

1.2 Organization of This Thesis

The organization of this thesis is as follows: In Chapter 2, we review researches on facial recall relating to forensic issues in the psychology literature and most of the facial recognition technologies used in the engineering literature. In Chapter 3, we introduce the forensic application software of facial recall, AICAMS-FIT. In Chapter 4, we describe the details of the component-based encoding scheme. Also, we report the performance and the shortcomings. In Chapter 5, we investigate the performance of PCA and LFA, in dealing with the problem of photo-to-sketch/sketch-to-photo matching. In Chapter 6, we suggest some future developments in this area. Finally, we conclude our thesis in Chapter 7.

Chapter 2

Literature Review

In this chapter, we review facial memory recall in forensic issues and facial recognition technologies in the field of psychology and engineering respectively.

2.1 Facial Memory Recall

Facial memory recall is one of the major topics of facial memory researches in the field of psychology. Although it is the most difficult practical problem for research on faces, it is necessary because of its application in forensic issues. For example, an effective system for recalling faces may help the police catch criminals. Researchers in this area usually relate their study to the forensic issues concerning the verbal and pictorial representation of a suspect face.

In [4], Ellis *et al.* have investigated the use of the Photo-Fit technique for recalling faces. They do two experiments to test the efficiency of the Photo-fit Kit for recalling faces. The first one is concerned with human ability to reconstruct photographs of faces made up of features randomly selected from the Kit. In the experiment, they examine two presentation procedures in which one has the face present and another

makes use of memory only. Their result indicates that people have difficulties in making up a reconstruction of a face using the Photo-fit Kit even when the original is present and all the necessary features are available. Although the performance of reconstruction of both procedures is far from perfect, they find that the face present condition produced better reconstruction than did the memory condition. The second experiment is concerned with the ease with which a photograph of a face could be identified from its Photo-fit reconstruction. Their result shows that the reconstructions of good encoders were more easily recognized than were those of poor encoder subjects. Finally, they conclude that they could not make any definitive recommendations concerning the use of the Photo-fit Kit by the police based on the present studies because the laboratory conditions are necessarily rather different to those existing in a real-life situation.

Later, in [5], M. A. Mauldin and K. R. Laughery study the effects of constructing an Identi-kit composite of a target face on subsequent recognition. Their results indicate that when subjects produce an Identi-kit composite of a target face, they are more likely to recognize the target face in a subsequent recognition task. However, they also show that generating verbal descriptions of features does not result in as much improvement in recognition as does the production process.

In [6], R. J. Phillips reports two exploratory studies on the relationship between recognition, recall and imagery of faces. Both studies suggest that subjects' reported imagery ratings for faces reflect performance on face recognition tasks more strongly than performance on recall tasks. He argues that the difficulty in recalling faces, no matter using verbal or pictorial description, is principally a problem of recoding information rather than retrieval. Finally, he thinks that an effective system of recall can only be achieved by training.

In [7], G. M. Davies has conducted a survey on the forensic facial recall. This survey reviews the principal methods used by the police for representing the facial appearance of a suspect. For examples, verbal descriptions such as *Cued Description* and *Prompted Description* or visual impressions such as the sketch written by police artist, the Identi-kit and the Photofit. It then reviews and discusses many researches on the relative success of visual and verbal procedures for transmitting likeness information and examines the influence of delay, sex and race on recognition. Based on the available evidence, the author concludes that there is no reason for suggesting that the visual medium is any more sensitive or efficient in conveying identity information than the verbal medium. Finally, the implications of the survey suggest a greater respect for accurate verbal descriptions as opposed to visual methods when interrogating witnesses.

2.2 Facial Recognition

Facial recognition is another topic of facial memory researches. However, here, we will focus on the study in the field of engineering. Researches on machine recognition of faces have been conducted over 25 years. During 1970's, typical pattern classification techniques using measured attributes between features in faces or face profiles were applied in facial recognition technology. During the 1980's, researches on facial recognition remained dominant. Since the early 1990's, research interest in face recognition technologies has grown significantly.

R. Chellappa, C. L. Wilson and S. Sirohey in [1] have conducted an excellent survey on human and machine recognition of faces. They not only report research activities in engineering literature, but also in psychophysics. In the survey, a lot of

facial recognition techniques and applications of still and range images are reported. These images include both front views and profiles of human faces. In their study, most of the face recognition technologies are approximately categorized into four approaches: Earlier Approach, Feature Matching, Neural Networks and Statistical Approaches. Here, we give a review on the literature of facial recognition based on the same categorization.

2.2.1 Earlier Approaches

Bledsoe in [8] proposes a system for computer recognition of faces. In this system, a human operator manually locates the feature points on the face and enters their positions into the computer. Given a set of feature point distances of an unknown person, nearest neighbor or other classification rules are used for identifying the test image. Since feature extraction is done manually, this system can handle wide variations in head rotation, tilt, image quality and contrast.

Y. Kaya and K. Kobayashi in [9] report a study on machine recognition of human face (in front view) using information theoretic arguments. In the study, they propose four necessary conditions for the selection of parameters characterizing a face. They suggest that the characteristic parameters should as easily be estimated from photographs as possible. Also, these parameters should not be greatly affected by any change of light condition and any change of facial expression. The least and the most essential one is that they should carry as much amount of information about a face as possible. The authors conduct an experiment to show that geometric distances between facial landmarks are some of effective characteristic parameters. In their experiment, sixty-two photographs were taken with special equipment to ensure consistent orientation and lighting conditions. Then they choose nine prominent

geometric distances to form a parameter vector. All of these geometric distances are measured by hand and normalized by the nose length. They construct a classifier based on the parameter vector and its estimate. This estimate is equal to the sum of the parameter vector and a distortion vector. The distortion vector consists of two components: one is the measurement noise and another is the inherent noise. Identification of a new face is done by comparing the threshold values, that are determined for each parameter from their statistical behavior, with the absolute norm between a stored parameter set and the test image parameter values. Since the parameters are fairly correlated, the actual dimension of the parameter vector may be reduced using Principal Component Analysis. The authors claim that they can get a probability of 92% of correct matching from 5000 faces when their algorithm is used.

2.2.2 Feature and Template Matching

R. Brunelli and T. Poggio in [10] develop a set of algorithms to assess the feasibility of recognition using a vector of geometrical features, such as nose width and length, mouth position and chin shape. Their geometric feature-based matching is loosely based on Kanade's work in [11]. Each facial images is preprocessed by setting the inter-ocular distance and the direction of the eye-to-eye axis to achieve scale and rotation invariance respectively. Translation invariance is achieved by setting the origin of coordinates to a point that can be detected with good accuracy. Eyes position is first determined using template matching by means of a normalized cross-correlation coefficient. Then geometric facial features are extracted by using the average anthropometric measures and applying the technique of integral projections. Horizontal gradients are used to detect the left and right boundaries of face and nose, while vertical gradients are used to detect the head top, eyes, nose base and mouth.

Facial outline is detected using dynamic programming which follow the outline on a gradient intensity map of an elliptical projection of the facial image. A total of 22 geometrical features are extracted as a feature vector for each face. Recognition is performed with a Nearest Neighbor classifier, with a suitably defined metric. Finally, a rejection threshold is introduced to enhance the robustness of classification.

Later, R. Brunelli and T. Poggio in [12] compare geometric feature-based matching with another matching strategy, the template matching. They set up an experiment with a database of 188 images, four for each of 47 people. Their model used in geometric feature-based matching is based on their previous works, while the one used in template matching is based on Baron's work [8]. The image is normalized using the same technique of geometric feature-based matching. Each person is represented by a database entry whose fields are a digital image of his/her frontal view and a set of four masks representing eyes, nose, mouth and face (the region from eyebrows downwards). The location of the four masks relative to the eye position is the same for the whole database. When attempting recognition, the unclassified image is compared with all of the database images, returning a vector of matching scores computed through normalized cross-correlation. The unknown person is then classified as the one giving the highest cumulative score. Recognition is, in fact, performed several times and single feature (mask) is used each time. The similarity scores achieved with different features are then added to obtain a global score. Their result shows that the use of template matching is superior in recognition performance. However, the feature-based strategy may allow a higher recognition speed and smaller memory requirements. The result is clearly specific to their task and to their implementation.

Ingemar J. Cox, Joumana Ghosn and Peter N. Yianilos in [13] introduce a mixture-distance technique for feature-based facial recognition in the setting where only a

single example of each face is available for training. As most of the work in facial recognition has concentrated on investigating alternative facial representations, their study pays more attention to the subsequent recognition algorithm. They focus on feature sets derived from the location of anatomical features in frontal or nearly frontal views of a face. They manually extract 35 points from each face and compute 30 according to the point measurement system introduced by J. Tojima and S. Sakamoto in [14]. All distances are normalized by the inter-iris distance to provide similarity invariance. Given a database of facial-feature vectors and a query consisting of a facial feature vector for some unidentified person, their objective is to locate the one in the database corresponding to the query. In their approach, a distance function is constructed based on local second order statistics as estimated by modeling the training data as a mixture of normal densities. They show that a flat mixture of mixtures performs as well as the best model and therefore represents an effective solution to the model selection problem. Their approach achieves a recognition rate of 95% on a database of 685 people, better than nearest neighbor search using Euclidean distance that yields only 84%. Finally, they conclude that the mixture-distances can be applied to more direct forms of the image ranging from raw pixels, through frequency transformations and the results of principal component and eigenface analyses.

2.2.3 Neural Network

M. Lades *et al.* in [15] present an object recognition system based on the Dynamic Link Architecture (DLA), which is an extension to classical Artificial Neural Networks. They choose the problem of face recognition to demonstrate the capabilities of the DLA. The DLA exploits correlations in the fine-scale temporal structure of cellular signals in order to group neurons dynamically into higher-order entities that

represent a very rich structure and can be used for encoding high level objects. A minimum of two levels, the image domain and the model domain, is needed for the DLA. The image domain corresponds to primary visual cortical areas and the model domain to the intertemporal cortex in biological vision. The image domain consists of a two-dimensional array of nodes, comprises different feature detector neurons which provide local descriptors of the image. Images are represented as attributed graphs. Attributes attached to the graph's nodes are activity vectors of local feature detectors, called "jets". Jets are based on Gabor-type wavelets. Neighboring nodes are connected by links, encoding information about the local topology. The model domain is an assemblage of attributed graph, being idealized copies of subgraphs in the image domain. There are excitatory connections between image domain and model domain. They are feature-type preserving, but not position specific. The DLA machinery is based on a data format able to encode information on attributes and links in the image domain and to transport that information to the model domain. The structure of signal is determined by three factors: the input image, random spontaneous excitation with neurons, and interaction with the cells of the same or neighboring nodes in the image domain. Binding between neurons is encoded in the form of temporal correlations. Four types of bindings are relevant to object recognition and representations: Binding all those nodes and cells together that belong to the same object, repressing neighborhood relationships within the image of the object, bundling individual feature cells between features present indifferent locations, and binding corresponding points in image graph and model graph to each other. The dynamic variable, "jet", plays the role of synaptic weights for signal transmission, controlled by the signal correlations between two neurons. Negative correlation value leads to a decrease and positive correlation value leads to an increase. Zero value leads to a resting state.

Recognition is based on elastic graph matching that compares stored model graphs with current image data in a two-stage optimization process, varying the image graph to minimize the cost of its match to the model. The cost function is a linear combination of an edge term and a vertex term, with a coefficient controls the rigidity of the image graph. During the first stage of optimization process, the image graph is shifted while keeping its form rigid. The shape of the model graph is initialized to an arbitrary position. During the second stage of the matching procedure the rigidity parameter is set to a finite value to permit small graph distortions and the vertices in the image domain can diffuse: they are visited sequentially and in random order and are shifted by a random vector below a preset maximum length. Each stage is terminated once a predefined number of trials have failed to improve the cost value or an optimal total cost is determined. Recognition takes place after the optimal total cost is determined for each object. The object with the best match to the image is determined. In the case of faces, if one face model matches significantly better than all competitor models, the face in the image is considered as recognized. L. Wiskott and C. v. d. Malsburg in [16] present a similar system for invariant and robust recognition of objects from camera images.

L. Wiskott *et al.* in [17] and [18] extend the ability of the preceding system in [16] in order to handle larger galleries and larger variations in pose and to increase the matching accuracy. The gallery may contain only one image per person. Three major extensions are made: They use the phase of the complex Gabor wavelet coefficients to achieve a more accurate node positioning and to disambiguate patterns that would be similar in their coefficient magnitudes. Also, they employ object-adapted graphs, so that nodes refer to specific facial landmarks, called fiducial points. The correct correspondences between two faces can then be found across large viewpoint

changes. They introduce a new data structure, called the bunch graph, which serves as a generalized representation of faces by combining jets of a small set of individual faces. The system is allowed to eliminate the need for matching each model graph individually and find the fiducial points in one matching process. Thus, the computational effort is reduced significantly. As jets are sensitive to phase rotation, they will have very different coefficients although they are taken from image points only a few pixels apart and represent almost the same local feature. In order to solve such severe problem for matching, the authors suggest either to ignore the phase or to compensate for its variation explicitly. They propose a general representation, rather than models of individual faces, to extract image graphs automatically for new faces. This representation is called a face bunch graph (FBG).

A FBG is a stack-like structure combining a representative set of individual model graphs and covering a wide range of possible variations in the appearance of faces, such as differently shaped eyes, mouths, or noses, different types of beards, variations due to sex, age, and race etc. A set of jets referring to one fiducial point is called a bunch. For example, an eye bunch may include jets from closed, open, female, and male eyes etc. in order to cover a wide range of local variations. The first set of FBGs is generated manually, then graphs for new images can be generated automatically by elastic bunch graph matching. The elastic bunch graph matching is based on graph similarity between an image graph and the FBG of identical pose. It depends on the jet similarities and the distortion of the image grid relative to the FBG grid. The matching process has two stages: In the first stage, the size and location of face is determined and the face image normalized in size. In the second stage, fiducial points are found and a precise image graph is extracted for recognition. The two stages use different FBGs with different emphasis and number of nodes. After extracting model

graphs from the gallery images and image graphs from the probe images, recognition is done by comparing an image graph to all model graphs and selecting the one with the highest similarity value, in a lower computational cost.

2.2.4 Statistical Approach

M. Turk and A. Pentland in [19] develop a system for facial detection and recognition using eigenfaces, that are in fact the eigenvectors (principal components) of the set of faces and face-like in appearance. These eigenfaces do not necessarily correspond to facial features such as eyes, ears, and nose, each one accounting for a different amount of the variation among the facial images. Eigenfaces corresponding to the highest eigenvalues are selected to form a face space. An individual face can be represented as a vector of weights by projecting itself into the face space through the operation of a simple inner product. When a test face is given for matching, it is also represented by its vector of weights. The recognition of the test image is done by locating the image in the database whose weights are the closest (in Euclidean distance) to the weights of the test image, provided that the test image is sufficiently close to the face space. The authors conducted several experiments to assess the performance of the approach under variations in lighting, size, head orientation, and selection of thresholds. Their results show that the approach is fairly robust to changes in lighting conditions, but degrades dramatically with size changes. Also, They extend their approach to real time recognition of a moving face image in a video sequence. The eigenfaces approach may also be implemented using a neural network architecture. For example, in a three-layer fully connected linear network, the input layer receives the input face image with one element per image pixel. The weights from the input layer to the hidden layer correspond to the eigenface, so that the value of each hidden

units is the dot product of the input image and the corresponding eigenfaces. The output layer produces the face space projection of the input image when the output weights also correspond to the eigenfaces.

Although Sirovich and Kirby in [20] demonstrate that a low-dimensional representation of faces in the lower dimensions of the face space (i.e., a representation that use as subset of eigenvectors with the largest eigenvalues) is optimal in the least-squares sense and hence minimizes the error of the reconstructed images, A. O'Toole *et al.* in [21] show that this representation is perhaps not the most useful one for recognition. They demonstrate that, in some cases, a various low dimensional representations of the faces in the higher dimensions of the face space (i.e. the dimensions associated with the eigenvectors with smaller eigenvalues) may provide better information for face recognition. They conclude intuitively that the eigenvectors with the larger eigenvalues are likely to convey information that is common to all the faces, whereas the eigenvectors with the smaller eigenvalues are likely to convey information about individual faces.

B. Moghaddam and A. Pentland in [22] and [23] extend the eigenface approach to view-based and modular eigenspaces for facial detection and recognition. These two extensions account for variations in head orientation, scale, hairstyle and makeup and lead to a more robust face recognition system. The view-based multiple-observer eigenspace technique generates a "view-based" set of separate eigenspaces, each capturing the variation of different individuals in a common view. This technique is compared with the parametric eigenspace method, which generates a universal eigenspace for all views. Their result shows that the view-based method is superior to the parametric method over a database of 189 images consisting of nine views of 21 people. The view-based method allows for recognition under varying head orientation

and scale. The modular eigenspace description technique generates separate eigenfaces for salient facial features, yielding eigenfeatures such as eigeneyes, eigennoses and eigenmouths. These eigenfeatures are incorporated with the eigenface to form a layered or modular representation of a face. This method achieves an asymptotic reocgnition rate of 98%. In both techniques, an effective indicator, called the distance from feature space, is used for robust detection and recognition.

B. Moghaddam, C. Nastar and A. Pentland in [24] introduce a novel approach for face recognition based on deformable intensity surfaces. The intensity surface of the image is modeled as a deformable 3D mesh in XYI space, which combines the spatial (XY) and grayscale (I) components of the image. The deformation process of the intensity surface takes place in 5 steps and it is ruled by a physically-based model, Lagrangian dynamics, which are in terms of the nodal displacements, the mass, the damping and the stiffness and the external force. The dynamics are solved for using the analytic modes of vibration. The nodal displacement is transformed into the modal amplitude subspace. The low-order strain energy, with high frequency components removed, is used as an efficient similarity measure. In [25], the modal amplitudes are further transformed a lower-dimensional subspace using the Principal Component Analysis.

In contrast to simple image similarity metrics methods such as Euclidean distance or normalized correlation, B. Moghaddam, W. Wahid and A. Pentland in [26] and [27] propose a similarity measure using a Bayesian analysis of image deformations for the deformable technique. They model two mutually exclusive classes of variation between two facial images: intra-personal and extra-personal. The intra-personal variation refers to variations in appearance of the same individual while the extra-personal variation refers to variations in appearance due to a difference in identity.

The similarity measure is expressed in terms of the *a posteriori* probability given by Bayes rule, using estimates of two likelihoods which are derived training data using an efficient subspace method for density estimation of high-dimensional data. The authors also apply this technique in [28] for eigenfaces technique. This time, the high-dimensional probability density functions for each respective class are obtained from training data using an eigenspace density estimation technique.

Due to the fact that much of the important information is contained in the high-order statistics of the images, M. S. Bartlett and T. J. Sejnowski in [29] and [30] present methods for representing facial images based on Independent component analysis (ICA) instead of PCA, which address the second-order statistics of the image set only. ICA is performed on the facial images under two different architectures: The first one provides a statistically independent basis set for the facial images that can be viewed as a set of independent facial features. The second one provides a factorial code for the facial images, in which the probability of any combination of feature can be obtained from the product of their individual probabilities. Their result shows that both ICA representations outperformed PCA representation, for recognizing images of faces sampled on a different day from the training images.

P. N. Belnumeur, J. P. Hespanha and D. J. Kriegman in [31] develop a face recognition algorithm which is insensitive to gross variation in lighting directions and facial expression. Their method is called Fisherfaces, which is a derivative of Fisher's Linear Discriminant(FLD). They project away variations in lighting and facial expression while maintaining discriminability, maximize the ratio of between-class scatter to that of within-class scatter. They conduct an experiment to compare the Fisher technique with other three methods: correlation, EigenFace and a variant of the linear subspace. Their results show that the Fisherface method achieved lower

error rates than the others.

The major problem in using PCA is that it invariably leads to global, nontopographic representations that are not amenable to further processing and are not biologically plausible. To solve the problem, P. S. Penev and J. J. Atick in [32] [33] propose a new mathematical construction called Local Feature Analysis for deriving local topographic representations for any class of objects. The representations of LFA are sparse-distributed and are effectively low-dimensional and retain all the advantages of the compact representations of the PCA. Unlike PCA, they give a description of objects in terms of statistically derived local features and their positions. The details of this algorithm as well as the PCA will be described in Chapter 4 and Chapter 5 respectively as our works depend on these two algorithms.

Chapter 3

A Forensic Application of Facial Recall

Generally, facial production or recall tasks can be categorized into verbal description and generating hard-copy representations [34]. There are a variety of techniques employed by law enforcement agencies for obtaining a visual representation of a target person's face. The three most widely employed and researched procedures are the sketch artist, Photofit and Identi-Kit. Each of these procedures involves the witness working with another person, an artist or technician/operator, to construct the face. The witness' task includes an onging verbal interaction with the artist or technician during which the face or parts of the face are being described. The Photofit and Identi-Kit involve the selection of individual facial features which are put together to form a composite face. Feature exchanges are then made to improve the match between the composite and the face in memory. This refinement process clearly includes a process of matching the current composite version to the face representation in memory.

3.1 Motivation

In Hong Kong, the police also employs the Identi-Kit for the facial recall tasks. When a witness performs the facial production task with a police artist or technician, he just selects facial features from the library of the transparent acetate sheets that match those of the target face in his memory. Then the police artist or technician combines the selected transparent acetate sheets to form a composite face by superimposing them in a special frame. Finally, the police artist portrays a sketch of the composite face and modifies it if necessary. The whole process is very time-consuming. In [35], the author points out that time is of the essence in forensic face memory because it may reduce memory by greater and greater amounts. He suggests that the use of computer may help to reduce the effect of time and provide an effective way in the generation of facial images from memory. Thus, an application software of facial recall called, AICAMS-FIT, is developed to achieve the above objective.

3.2 AICAMS-FIT

AICAMS-FIT stands for Artificial Intelligence Crime Analysis And Management System - Facial Identi-kit. It is one of the components of a joint project conducted by the Chinese University of Hong Kong and the Hong Kong Police. Its main purpose, as mentioned above, is to replace the time-consuming manual process of facial recall using transparent acetate sheets in the past and speed up the current process of facial recall. AICAMS-FIT has been developed for more than two years. During the development, we keep close contact with the Police so that the application fits the use in the forensic agency. Improvements have been made continuously to make the application benefit the process of facial recall more and more. Recently, there are two

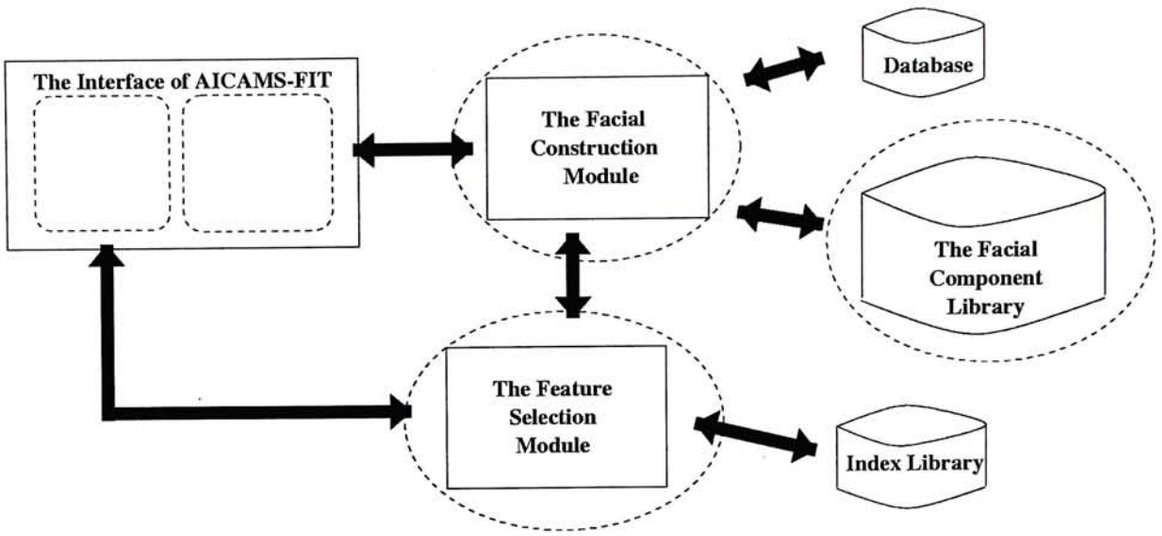


Figure 3.1: The General Architecture of AICAMS-FIT.

different versions of AICAMS-FIT: the operation version and the research version. The operation version is now deployed in the Police Department while the research one is the improvement of the operation one and still in progress. However, in this thesis, we will not discuss the difference between these two versions. Despite the difference in software versions, the general architecture of AICAMS-FIT is the same and is shown in Figure 3.1. Although the figure shows several components in the system, it is mainly made up of three components: the Facial Component Library, the Feature Selection Module and the Facial Construction Module. We will describe the details of the three components in this section.

3.2.1 The Facial Component Library

AICAMS-FIT provides twelve classes of facial features stored in its Facial Component Library. Six of them are the essential feature classes that compose a complete human face. They are hair including the forehead, facial outline including the ears,

eyebrows, eyes, nose and mouth. Almost every people will have these classes of facial features. Also, they are comparatively invariant (except the hairs). Table 3.1 shows the quantity of facial features in these six major classes. Based on this quantity, we can generate at least 1.42×10^{12} different composite faces. The remaining six feature classes contain facial features that belong to some particular groups of people. For examples, male has beard and moustache, old people have wrinkles and sidelines of nose and short-sighted people wear spectacles. Table 3.2 shows the details of these six minor classes in which the class of "Others" contains facial features such as scars, moles and acne. In our research, we mainly focus on the six essential feature classes.

All the facial features in the library are extracted from a large set of photographs of local citizens by a professional forensic artist based on his expertise. Most of them are extracted from male faces because, in forensic agencies such as police, most of their target people are male. However, this does not imply that they cannot be used to construct a female face. The artist then portrays the features and we store them as feature templates[1] with alignment i.e. each facial feature is drawn at the center of the template. The size of the template is fixed for each feature class. For example, the size of nose template is 90×65 pixels and the size of the mouth template is 60×100 pixels. He further sub-classifies the feature templates of each feature class into different types manually. For examples, eyes can be sub-classified into round eyes, slit eyes, slim eyes, triangular eyes and etc. The sub-classification of the six essential feature classes is shown in Table 3.3. Moreover, features are indexed within its feature class. That is, for each feature class, features are assigned an integer so that they are arranged in order. The user can search the required facial feature by the index.

[1] A template here can be interpreted as a picture.

Facial Feature	Hair	Facial Outline	Eyes	Eyebrows	Nose	Mouth
Quantity	94	130	123	112	82	103

Table 3.1: The major facial features in AICAMS-FIT.

Facial Features	Nose Sidelines	Moustaches	Beard	Wrinkles	Spectacles	Others (e.g scars)
Quantity	7	17	12	13	58	12

Table 3.2: The minor facial features in AICAMS-FIT.

Feature Class	Sub-Classes
Hair	Short & Straight, Short & Curly, Long & Straight, Long & Curly, Bald
Face Outline	Triangular, Long, Square, Round
Eyes	Round, Slit, Slim, Triangular, Common
Eyebrows	Bushy, Thin, Narrow, Common
Nose	Small, Large, Pointed, Common
Mouth	Thick, Thin, Protruding Teeth, Common

Table 3.3: Sub-classification of the six essential feature classes.

3.2.2 The Feature Selection Module

Figure 3.2 shows the user interface of AICAMS-FIT. The Feature Selection Module is shown on the left-hand side of the figure. This module provides a friendly interface for the user to select the desired facial features that match the one in the user's memory. Each time, the user can browse a set of facial features and he can change to another set until the desired feature is seen. Also, it allows the user to constrain the set of facial features for selection by choosing the desired subclasses of feature class. For example, when the user wants to select the hair, he can limit the set into long-type or short-type only.

The facial features shown in the Feature Selection Module are actually taken from the Index Library. The Index Library is in fact a duplication of the Facial Component Library but the feature templates stored in it are processed in a much coarser resolution so that it requires less memory storage. The need of the Index Library arises from the following consideration: The feature templates in the Facial Component Library are so large that only few of them are allowed to be shown on the screen, especially those of the hair. However, for browsing purpose, we should allow more and more features for the user to choose. Therefore, smaller versions of the templates are needed. The Index Library has the same indexing method as the Facial Component Library.

3.2.3 The Facial Construction Module

The right-hand side of Figure 3.2 shows the Facial Construction Module. This module displays the selected facial features and combines them to form a composite face using a layer-by-layer representation and an overlay algorithm as shown in Figure 3.3. In the layer-by-layer representation, each layer contains one feature template only. The

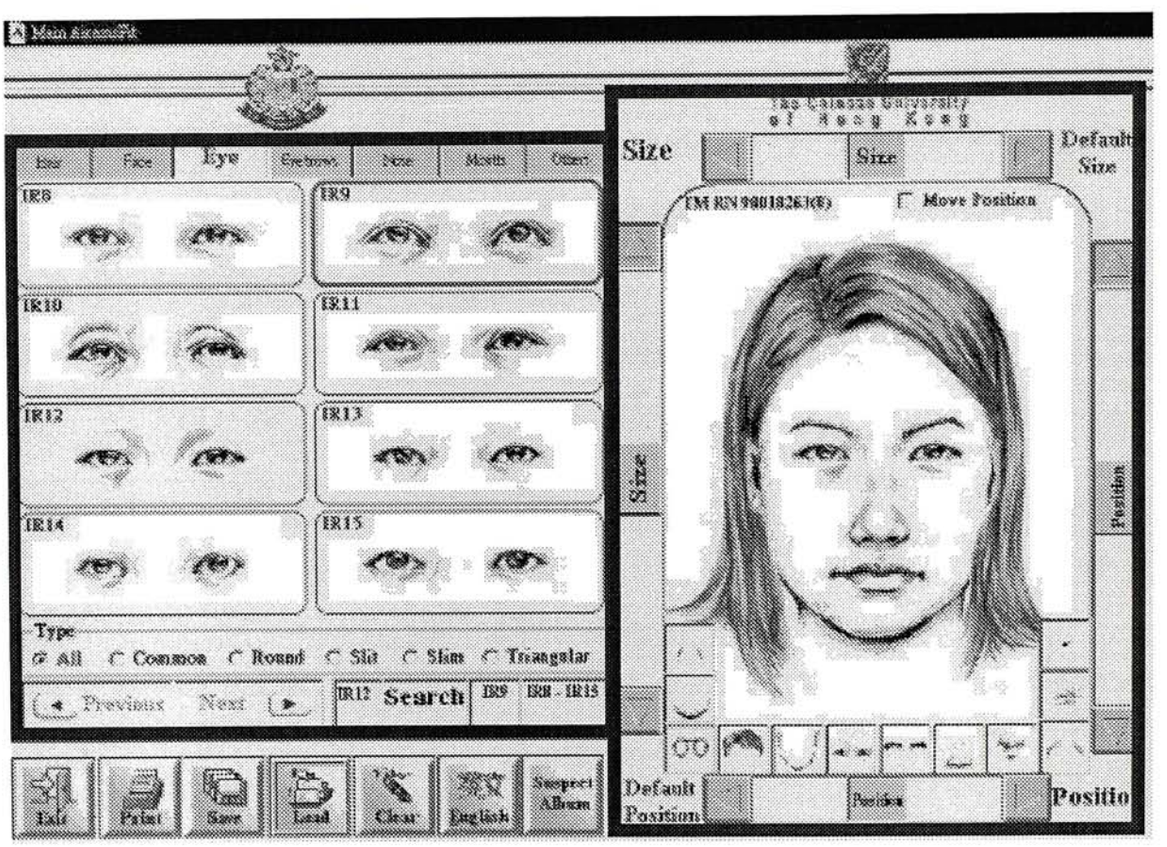


Figure 3.2: The User Interface of AICAMS-FIT.

The area enclosed by the solid dark square on the left-hand side indicates the feature selection module while the area enclosed by the solid dark rectangle on the right-hand side indicates the facial construction module.

placement of layers obeys the following top-down order: Spectacles (S) > Hair (H) > Beard (B) > Eyebrows (E) > Others (O) > Nose (N) > Moustaches (M) > Mouth (U) > Eyes (I) > Wrinkles (W) > Nose Sidelines (NS) > Facial Outline (FO). The symbol >, here, is used to represent the relation: If $A > B$, then A is on top of B. The symbols in the brackets are used as the short forms of facial features for convenience. This top-down order also represents the priority of overlay that is used to determine which layer can overlie the others. We can see that when feature templates are placed layer-by-layer, there are inevitably overlapping areas between them. According to our priority of overlay, the templates in upper layers will overlie those in the lower layers. For example, given two templates, H and I, since $H > I$, $H \cap I \in H$. That is, we will see the part of hair instead of the part of eyes within the overlapping area. Formally, the overlay algorithm is expressed as follows:

Given a set of facial features $\{F_i, i = 1, 2, \ldots, m\}$, where m is the total number of features,

$\forall j, k \in \{1, 2, \ldots, m\}, j \neq k$

If layer(F_j) > layer(F_k) then

overlay(F_j) > overlay(F_k)

$F_j \cap F_k \in F_j$

else overlay(F_j) < overlay(F_k)

$F_j \cap F_k \in F_k$

where layer($\bullet$) denotes the layer of the feature and overlay($\bullet$) denotes the priority of overlay of the feature.

However, this will raise the problem that the major part (the picture of the feature) of a template in the lower layer may be overlaid by the unnecessary parts (the background) of those in the upper layers. To overcome this problem, we have made

the background (the part in white) of all feature templates transparent.

Since the feature template is portrayed by the artist, the composite face is in fact a composite of images of sketch instead of photographic images. Actually, once a facial template is selected, it will be placed at a default position. The default position is determined from the standard human face shown in Figure 3.4, which is used as the reference by the artists, and obeys the topography of a human face. Also, the module allows the user to vary the size and the location of each selected facial feature horizontally or vertically in order to make the composite face more similar to the face in the user's memory. The size variation uses a linear approach such that the width w and the height h of the feature template are enlarged or reduced by a constant factor at time t as follows:

$$w_{t+1} = w_t + \alpha_i \tag{3.1}$$

$$h_{t+1} = h_t + \beta_i \tag{3.2}$$

where α_i and β_i are the width constant and the height constant for feature class i respectively. Different feature classes have different width constants and height constants.

Once a composite face is constructed, it is represented as a set of tuples internally. When the user wants to store the constructed face, the set of tuples is used instead of the templates in the database. In the mathematical term, the set of tuples is represented as $\{(F^i_{index}, F^i_x, F^i_y, F^i_w, F^i_h), i = 1, 2, \ldots, m\}$, where m is the total number of feature classes, $F^i_{index}, F^i_x, F^i_y, F^i_w$ and F^i_h indicate the index, the x-coordinate, the y-coordinate, the width and the height of the selected feature template of class i respectively. Hence, we can see that a composite face is encoded as a set of numbers.

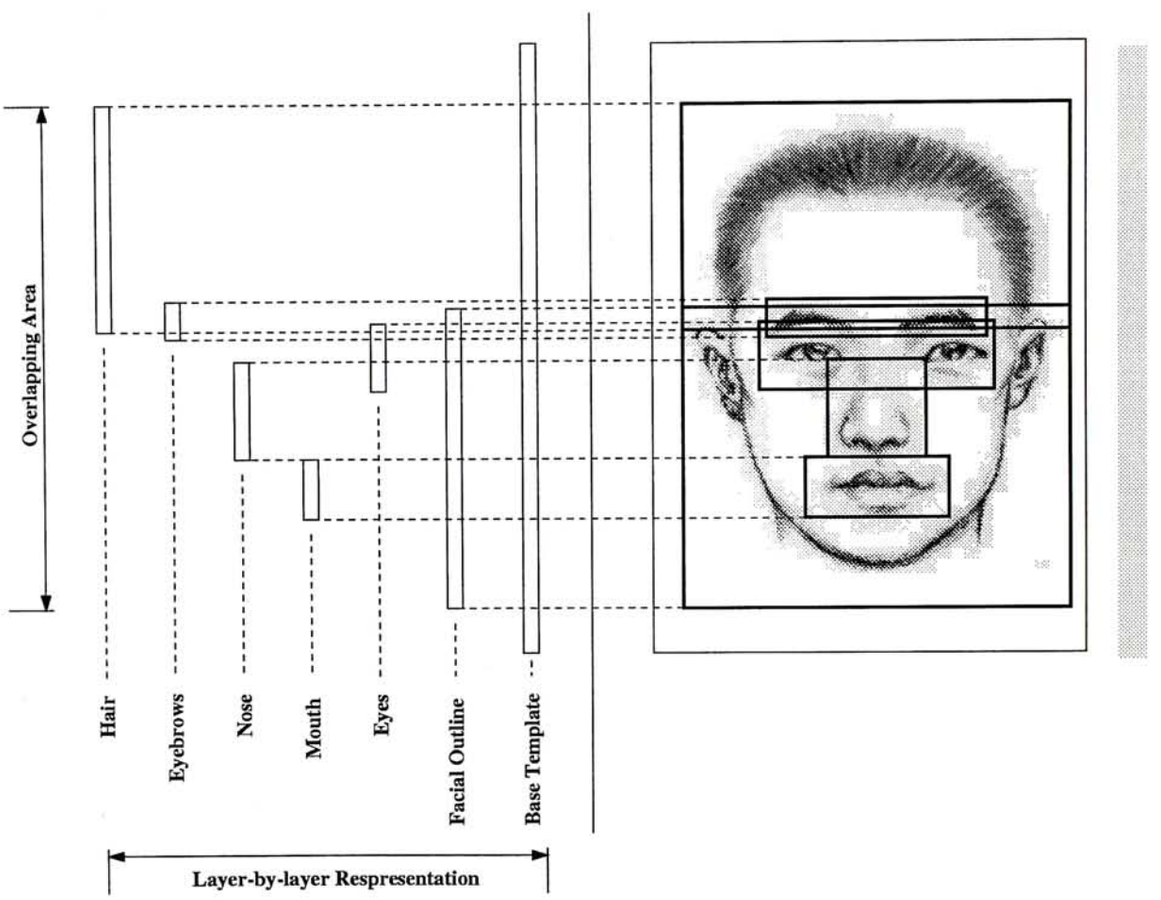


Figure 3.3: The production of a composite face.

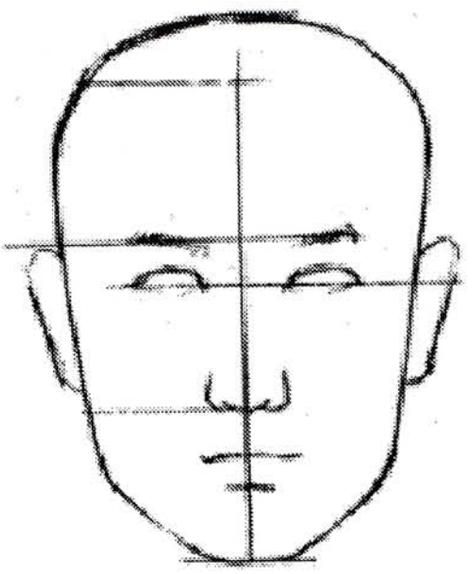


Figure 3.4: The standard human face.

3.3 The Interaction Between The Three Main Components

The interaction between the Feature Component Library, the Feature Selection Module and the Facial Construction Module is described as follow: Once the user selects a facial feature from the display of the Feature Selection Module, the index of such feature will then be sent to the Facial Construction Module. It will retrieve the corresponding facial feature from the Feature Component Library according to the index. A composite face can be seen if all the necessary features are chosen. The user can modify the existing composite face using the facilities provided by the Facial Construction Module. Finally, the user can store the required composite face, as a set of numbers instead of a set of feature templates, into the database.

3.4 Summary

In this chapter, we have presented an application of facial recall called AICAMS-FIT that is used in Hong Kong Police. The system is developed to speed up the process of facial recall in forensic issues. It is composed of three main components: The Feature Component Library, The Feature Selection Module and The Facial Construction Module. The output of the system is a composite face which is made up of feature templates in sketch format. The user can easily create such a composite face through a friendly user-interface.

Chapter 4

Sketch-to-Sketch Matching

The forensic tasks of facial memory can be categorized into two parts: facial production (recall) and facial matching (recognition) [34]. Actually, facial matching is always the subsequent procedure of facial production. In this chapter, we propose a component-based encoding scheme for AICAMS-FIT to perform facial matching in an efficient way. By the encoding scheme, each class of facial features has it's own similarity matrix. Facial matching is simply done using the similarity matrices. Since there are limitations of size variation and position variation, We will investigate their effect on facial matching.

4.1 The Representation of A Composite Face

In facial recall, a complete face is usually produced as a composite face which consists of at least the six essential features as the one described in Chapter 3. In terms of mathematical formulation, we can express this kind of composite face as $C = \{P_{i,k}, i = 1, 2, \ldots, m\}$, where m is the total number of feature classes, $k \in \{1, 2, \ldots, n_i\}$, n_i is the total number of facial features in feature class i and, $P_{i,k}$ indicates that the facial

feature k of feature class i is selected.

4.2 The Component-based Encoding Scheme

Facial matching can be regarded as a matching of individual facial features. Traditional facial matching requires real time comparison of facial features. This is true either in feature-based matching or template-based matching because we cannot determine the similarity between facial features of two different faces prior to the matching process. However, if a face is made up of facial features from a facial feature library, such as the composite face constructed in the application of facial recall, AICAMS-FIT, batch mode comparison of facial features can be done. That is, we can pre-compare the facial features prior to the facial matching. When we want to compare two composite faces, we can just make use of the pre-computed values. Time can be greatly reduced for facial matching if the real-time comparison requires intensive computation.

The component-based encoding scheme is proposed to achieve this objective. It encodes the similarity between each facial feature of the same feature class into a similarity matrix, so that the matching between two composite faces is made simpler. It can be as simplest as summing up the similarity value obtained from the similarity matrix of each feature class. Although it requires quite a lot of work in the preprocessing stage, it favors later process of facial matching. The idea of the encoding scheme is simple and intuitive. It is described as follows. Suppose there are two composite faces C_1 and C_2, where $C_1 = \{P_{i,k1}, i = 1, 2, \ldots, m\}$ and $C_2 = \{P_{i,k2}, i = 1, 2, \ldots, m\}$. They are nearly identical and differ only in the facial features of the feature class j, i.e. when $i = j$, $k1 \neq k2$. Hence, a measure of sim-

ilarity between C_1 and C_2 will only take account of the similarity between $P_{j,k1}$ and $P_{j,k2}$.

$$Sim(P_{j,k1}, P_{j,k2}) = Sim(C_{j,k1}, C_{j,k2}) \tag{4.1}$$

where all facial features of $C_{j,k1}$ and $C_{j,k2}$ are identical except the one in feature class j. The problem here is how to find out $Sim(C_{j,k1}, C_{j,k2})$, i.e. we have to deal with the problem of sketch-to-sketch matching of faces.

In the literature, researchers always focus on the recognition of photographic images. The face recognition technologies mentioned in Chapter 2 are all proposed for such matching either the faces are in frontal view or profile. Seldom researches concern the recognition of images of sketch. Although the authors in [36] compare facial line drawings with photographs, the situation is different because we want to compare a facial sketch with other sketches. However, when we consider an object, the main difference between its image of sketch and its photographic image is only the intensity distribution. In [37], Pearson and Robinson describe an automatic method for producing computer-drawn facial sketches which appear very similar to those produced by a human artist. Their cartoon algorithm comprises two components: one draws lines at the locations of intensity changes corresponding to luminance valleys and edges (the 'valledge' detector) while another applies a 'threshold' to the original intensity distribution, and replaces any area darker than the threshold with black. Hence, we can treat an image of sketch as a transformation from a normal photographic image. Since we are now concerning sketch-to-sketch matching, both the test image and the searching images are sketches. We can assume that they are under the same transformation of their corresponding photographic faces although actually those photographic faces may not exist. Therefore, we may try to use any modern face recognition technologies to compare two images of facial sketches in order to get

$Sim(C_{j,k1}, C_{j,k2})$. In our research, we have chosen the Local Feature Analysis [32], [33] because it is quite a robust algorithm in solving the problem of facial matching of photographic images.

4.2.1 Local Feature Analysis

Local Feature Analysis (LFA) is a two-stage procedure for constructing a local topographic representation of objects in terms of local features from the global Principal Component Analysis (PCA) modes[1].

In the first stage, LFA derives a dense set of local feed-forward receptive fields, defined at each point of receptor grid and different from each other, that is optimally matched to the input ensemble, and whose outputs are as decorrelated as possible. Suppose there is an input ensemble of sensory signal denoted by $\{\phi^t(x), t = 1, \cdots, T\}$ where T is the total number of examples in the ensemble and $\phi(x)$ is a sensory signal with V total sampling points that possess some topography (Since we are dealing with images, $\phi(x) = I(x)$ with x the 2D grid of photoreceptors), then the LFA outputs are defined as

$$O(x) \equiv \int K(x,y)\phi(y) = \sum_{r=1}^{N} \frac{A_r}{\sqrt{\lambda_r(x)}} \Psi_r(x) \tag{4.2}$$

where A_r is the PCA coefficient, $\Psi_r(x)$ is the PCA eigenmode with $\lambda_r(x)$ being its corresponding eigenvalue and $N < T \ll V$. Also

$$K(x,y) = \sum_{r=1}^{N} \Psi_r(x) \frac{1}{\sqrt{\lambda_r}} \Psi_r(y) \tag{4.3}$$

$$P(x,y) = \sum_{r=1}^{N} \Psi_r(x)\Psi_r(y) \equiv \langle O(x)O(y) \rangle \tag{4.4}$$

$K(x,y)$ is the kernel of the representation, and $P(x,y)$ is the residual correlation of the outputs. Topography is maintained here as the kernels of the representation are

[1]PCA is defined in Section 5.1.

labeled with the grid variable x instead of the PCA eigenmode index r.

An example $\phi(x)$ can be reconstructed by

$$\phi^{rec}(x) = \int K^{(-1)}(x,y)O(y) \tag{4.5}$$

where $K^{(-1)}(x,y) = \sum_{r=1}^{N} \Psi_r(x)\sqrt{\lambda_r}\Psi_r(y)$. However, as V outputs ,$O(x)$, are needed, the dimensionality is not reduced.

Since the objects from the input ensemble span only a very low dimensional subspace, the dense outputs are necessarily linearly dependent and also, contain residual correlation; this temporarily obscures the low dimensionality of the ensemble. In the second stage, the residual correlations $P(x,y)$ are used to sparsify the output in order to create a local sparse-distributed, explicitly low dimensional representation. This representation requires only a small number of active outputs for any given input. The number of active units is on the order of the PCA dimensionlity, and the location of the activity changes from one sensory input to another, thereby signaling explicitly additional information about the locations of the current object's features. Serial sparsification algorithm is applied to obtain this final representation. The algorithm is described as follows:

Starting with the empty set $M^{(0)} = \emptyset$ and add a point to M, chosen according to the following criterion. At the n-th step, three things have to be done.

i. Given the current set $M = M^{(m)}$, calculate the current maximum likelihood estimate (m.s.e) on $O(x)$ by

$$O^{rec}(x) = \sum_{m=1}^{|M|} a_m(x)O(x_m) \tag{4.6}$$

and calculate the current error by

$$E = \langle \| O^{err}(x) \|^2 \rangle \equiv \langle \| O(x) - O^{rec}(x) \|^2 \rangle \tag{4.7}$$

where the optimal linear prediction coefficients $a_m(x) = \sum_{l=1}^{|M|} P(x, x_l)(P^{l-1})_{lm}$ with $P(x_l, x_m) \equiv P'_{lm}$.

ii. Seek for the grid point x_{m+1} that has the maximum value of the reconstruction error $O^{err}(x_{m+1})$. If E is below some acceptable level, the sparsification is terminated.

iii. If it is not terminated, add x_{m+1} to $M^{(n)}$, which results in $M^{(n+1)}$ and go to step $n+1$ until $N = |M|$ points are selected. When $N = |M|$, the entire $O(x)$ is reconstructed without error.

The representation produced by the sparsification algorithm reveals the low dimensionality of the object space.

Recently, the LFA algorithm is implemented in a commercial facial recognition engine called FaceIt[2] and we make use of this engine to obtain $Sim(C_{j,k1}, C_{j,k2})$.

4.2.2 Similarity Matrix

The construction of similarity matrices in our encoding scheme is described as follows: Firstly, for each class of facial features, e.g. the mouth, we generate a set of identical artificial human faces [3] in which only the facial features of that class are different. It is used as the searching database. Then we choose one of the composite faces in the searching database as a test face. We use the FaceIt Engine to compute the similarity between the test face and the searching database. Then we get the similarity scores that indicate how a facial feature of a feature class is similar to the others of the same class. The similarity score (confidence level) obtained by the engine is between 0 and

[2]FaceIt is a registered trademark of VISIONICS CORPORATION

[3]We construct these faces without any reference of real people. However, since the facial features are portrayed from real people, they look like human faces perceptually.

100. The process is duplicated until all facial features have been matched. Figure 4.1 shows the whole idea of the encoding scheme.

Actually, the set of artificial faces, for each class of facial features, is generated by varying the choice of facial feature of that class and fixing the choice of facial features of other feature classes. The choice of other features is determined in a random fashion. In our work, we use only one set of such faces based on one fixed choice of other features to construct the similarity matrix of one feature class because we assume that the similarity between facial features of the same feature class is independent of the choice of other features.

Once we get all the similarity scores of a feature class, we do normalization before putting them into the similarity matrix of that feature class. Suppose we get a set of similarity scores for feature class j, $S = \{Sim(P_{j,ki}, P_{j,kl}), i = 1, 2, \ldots, n_j, l = 1, 2, \ldots, n_j\}$, then they are normalized by,

$$s_i^N = \frac{s_i - s_{min}}{s_{max} - s_{min}} \tag{4.8}$$

where $s_i \in S$, s_{min} and s_{max} for class j are obtained from $max(s_i)$ and $min(s_i)$ respectively. See Figure 4.2. After normalization, the similarity scores are finally rescaled into values between 0 and 1. Then we can get a $n_j \times n_j$ similarity matrix for feature class j. Since the similarity scores obtained from FaceIt Engine are commutative, i.e. $Sim(P_{j,ki}, P_{j,kl}) = Sim(P_{j,kl}, P_{j,ki})$, the similarity matrix is in fact symmetric. Thus, the storage can be reduced.

Since we have to build the similarity matrices for the six essential feature classes, we prepare six sets of composite faces in which each set contains identical composite faces except facial features of one feature class are different. For example, in SET 1, we have a set of composite faces with different hairs but other features are the same; In SET 2, we have a set of composite faces with different facial outlines but other

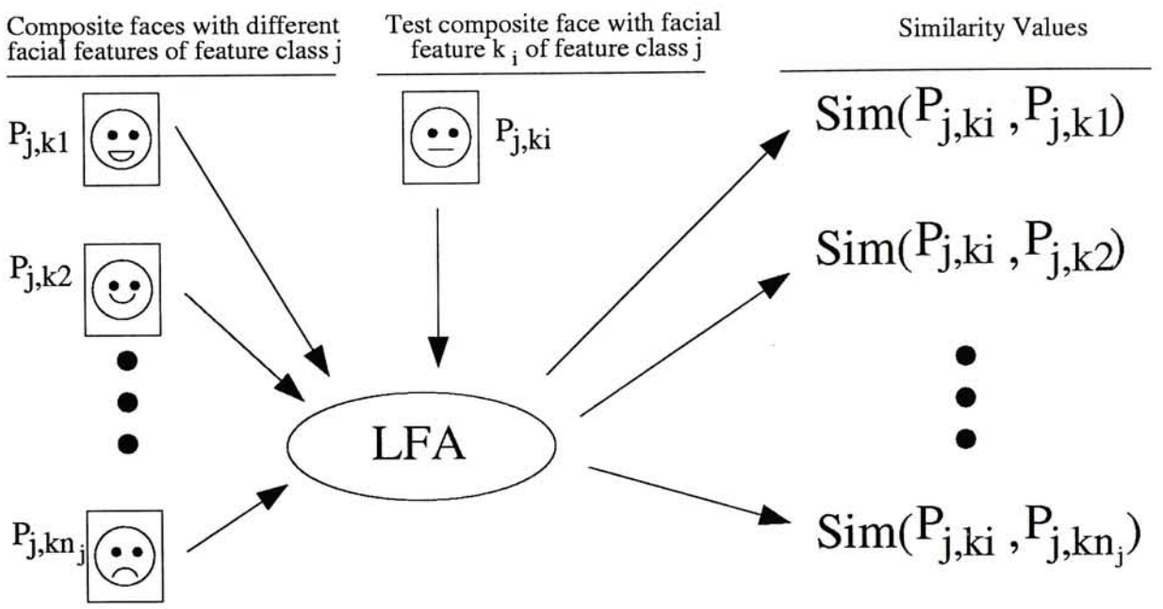


Figure 4.1: The component-based encoding scheme using LFA.

Identical composite faces with different facial features of feature class j.

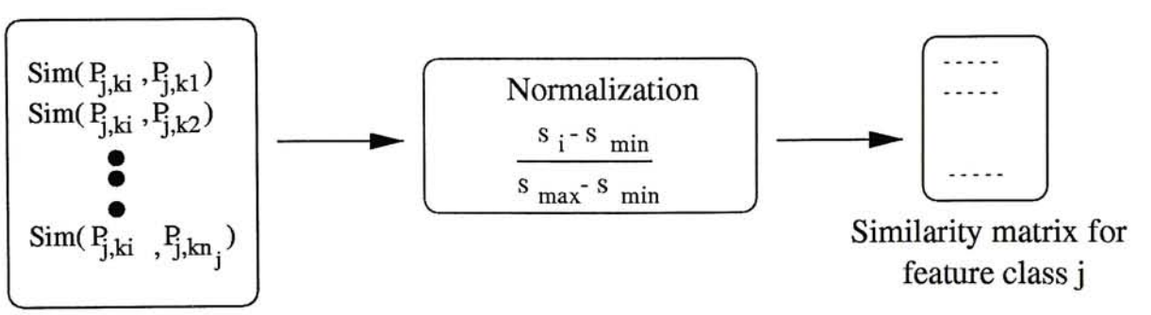


Figure 4.2: The construction of a similarity matrix.

features are the same; and so on. Sample of composite faces of the six essential sets are shown in Figure 4.3.

After constructing the similarity matrices for the six feature classes, we can now compare two composite faces, C_x and C_y, through the following similarity measure:

$$Sim(C_x, C_y) = \frac{1}{m} \sum_{j=1}^{m} Sim(P_{j,ki}, P_{j,kl}) \tag{4.9}$$

It is, in fact, the average of similarity values obtained from the six similarity matrices. The range of the value is between 0 and 100. Facial matching is performed when a novel composite face is compared with those in the database. Every composite face in the database will be ranked according to the result of the similarity measure and the one with the top rank is regarded as the most similar one. Figure 4.4 shows a sample case of matching a composite face (the left one) with a database of 80 faces.[4] In this example, we can see that the top ranked one shown on the right looks like the test face.

In [38], the authors point out that, in the area of feature selection, some discrete features such as the eyes, mouth, chin and nose have been found to be important cues for the discrimination and recognition of faces. In other words, different facial features may have different discrimination power on facial matching. Therefore, we add a weighting factor for each of the facial features and equation 4.9 now becomes:

$$Sim(C_x, C_y) = \sum_{j=1}^{m} \omega_j \times Sim(P_{j,ki}, P_{j,kl}) \tag{4.10}$$

where $\sum_{j=1}^{m} \omega_j = 1$. In the current system, the weighting factors are adjusted by the user.

An advantage of the encoding scheme is that it allows facial matching using only a few of facial features as the matching criterion. Figure 4.5 shows such kind of

[4]In the figure, the number of faces in the database is 81 because it includes the test one.

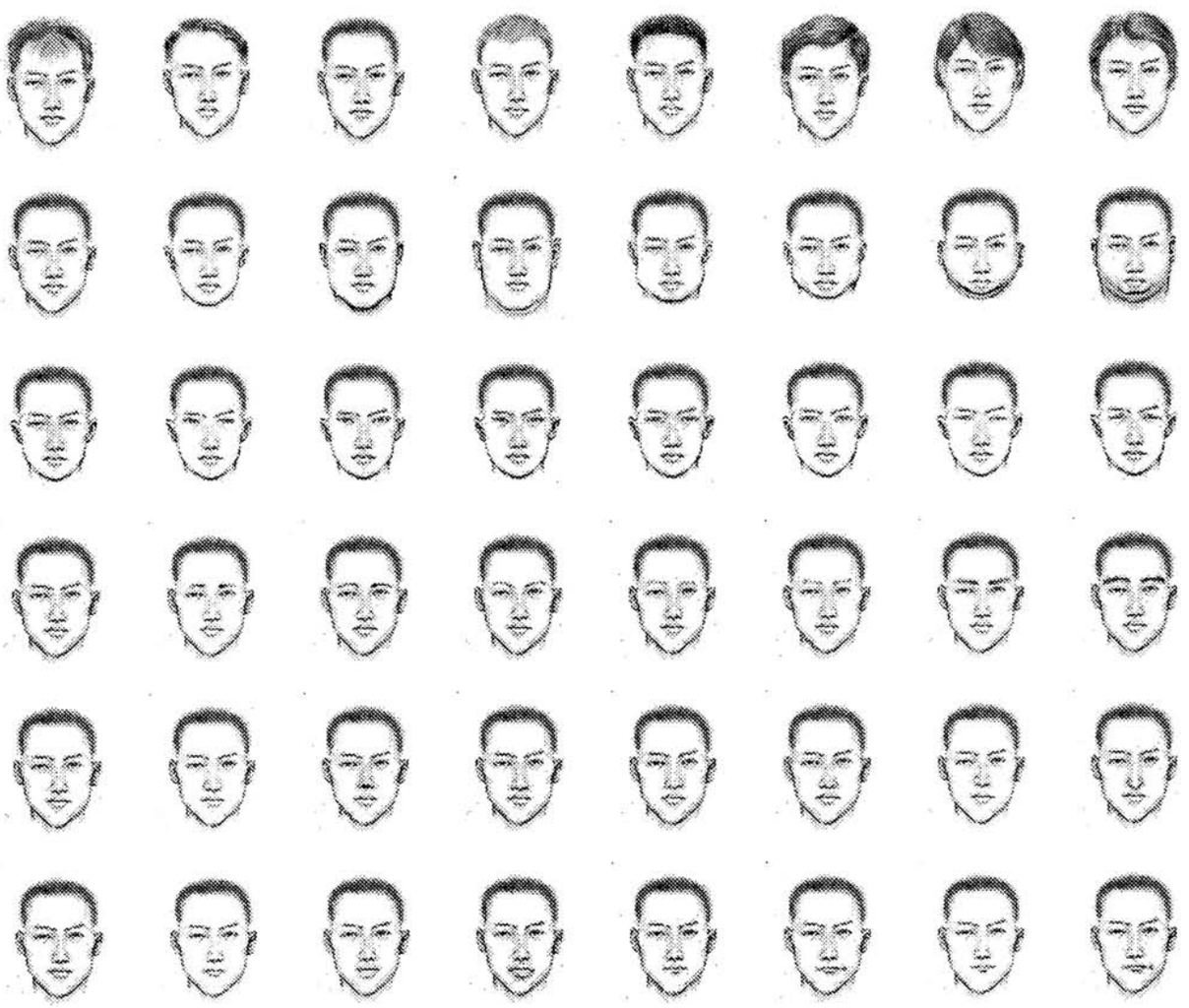


Figure 4.3: Sample of identical faces with only one different facial feature. The first row shows part of identical faces with different hairs; The second with different facial outlines; The third with different eyes; The fourth with different eyebrows; The fifth with different noses and the last with different mouths.

matching: we perform matching based on the hair and eyes only, i.e. we choose $\omega = 0.5$ for both hair and eyes while others are zero. The result shows that the first four top ranked faces on the left have the same hair and similar eyes as the test one on the right. In real situation, witness may remember only one or few salient facial features of the target person, so other essential features may be missed in the reconstructed face. However, it is also allowed to perform matching with such incomplete face under our encoding scheme because we can select the facial features we wanted. Moreover, in crime cases, suspects may disguise in various ways. For example, they may wear eyeglasses or stick a false-moustache onto their faces. Figure 4.6 shows an example: a face with eyeglasses is matched with a database of 80 faces. Since our encoding scheme ignores the effect of such kind of facial features on facial matching, the result matches our expectation that the resulting face ranked as the top shown on the right looks similar to the test face if it is without eyeglasses.

4.3 Experimental Results and Evaluation

In the traditional evaluation of face recognition technologies, it usually requires multiple examples of faces for each human object so that each face in the test set will have its own corresponding face in the searching set as they belong to the same person. If the retrieved top ranked face belongs to the same person as that of the test face, it is considered as a hit. Performance is evaluated by considering the hit rate. However, in the application of facial recall, the evaluation is difficult because it usually has only one single recalled composite face for each human object. And its subsequent matching requires only to find out a face in the database that is most similar to the test face no matter the test face is identified to belong to someone. Therefore, in our

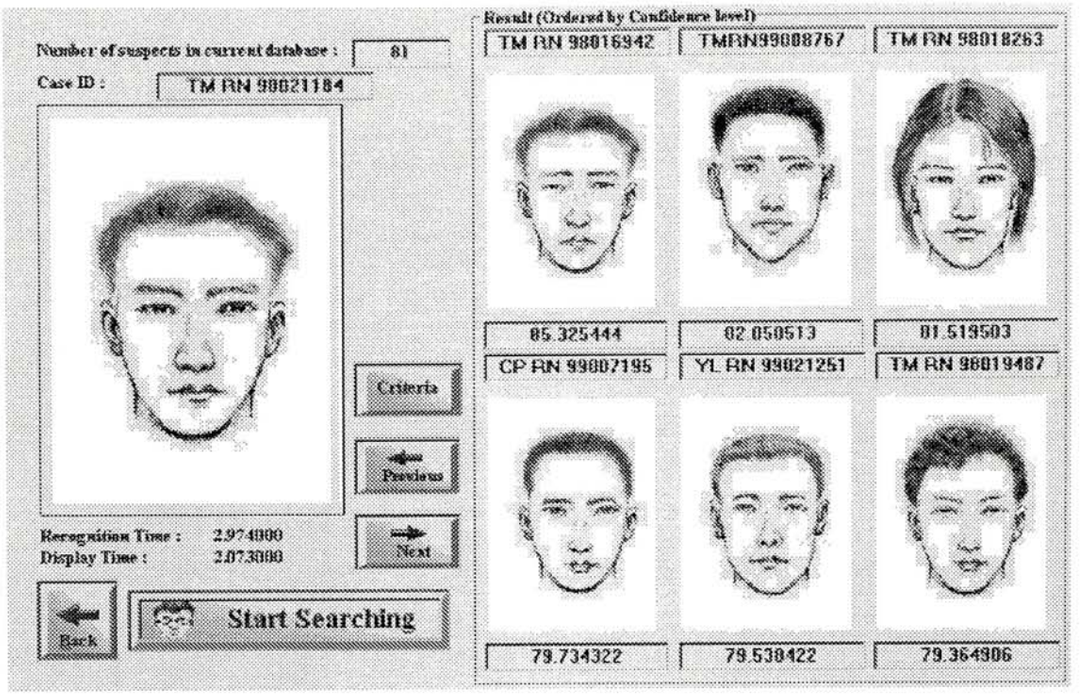


Figure 4.4: Sample result of sketch-to-sketch matching using component-based encoding scheme.

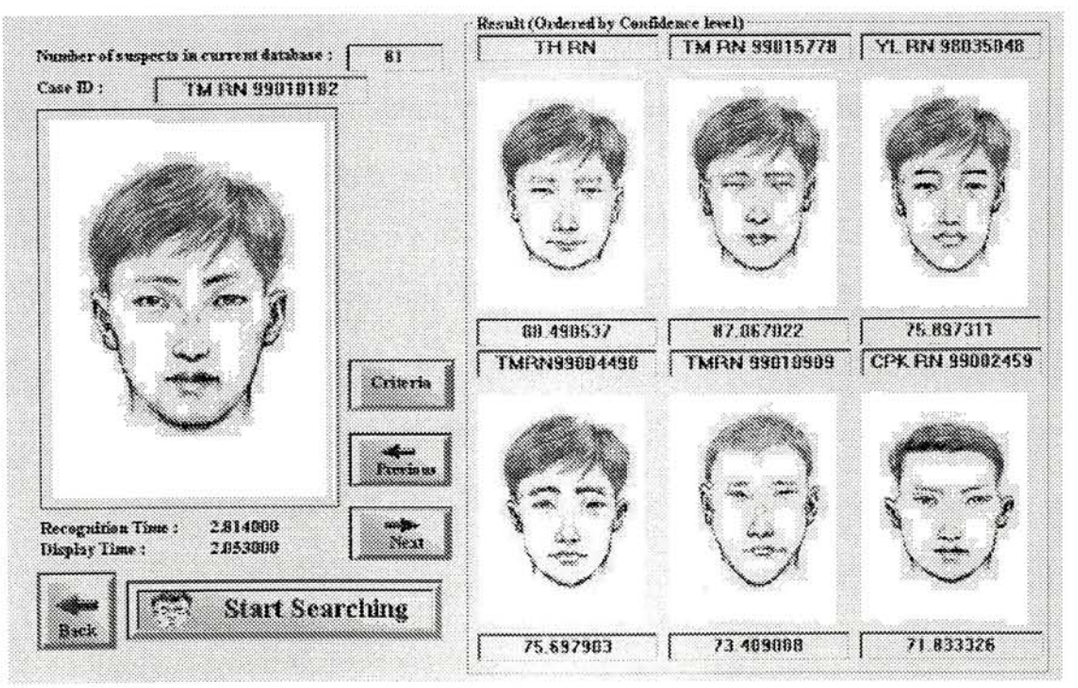


Figure 4.5: An example of matching based on fewer facial features.

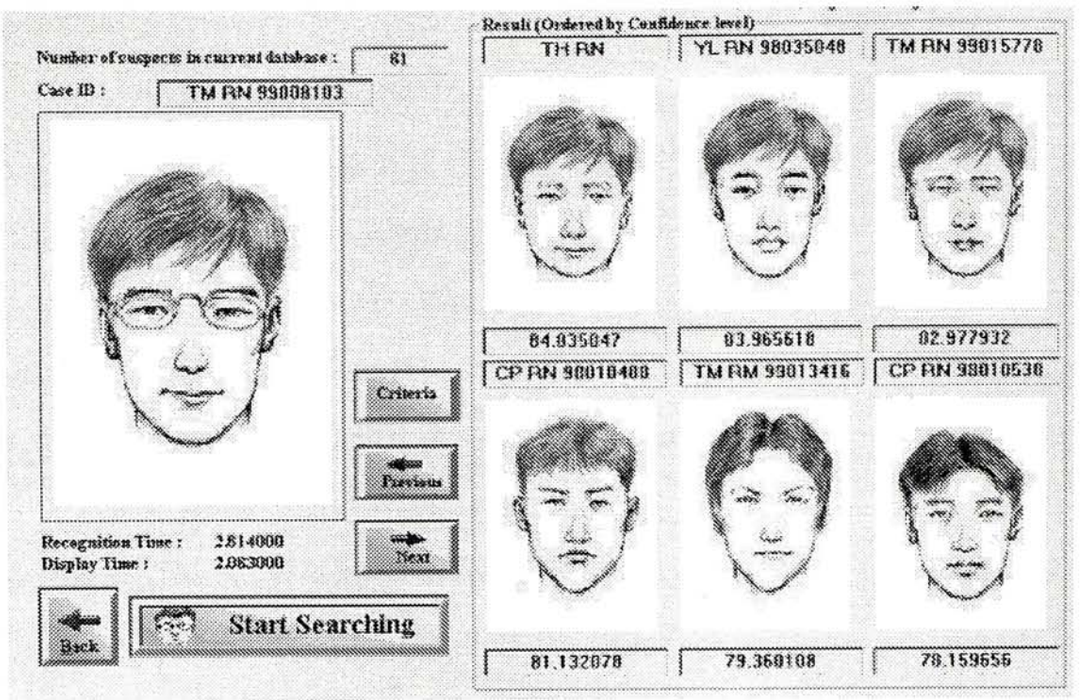


Figure 4.6: An example of matching based on minor facial features.

experiment, we require human judgement to find out the target face in the database for each test face for our evaluation. We have collected a database of 247 composite faces from the police. Since not all of the composite faces are suitable [5], we sample only 100 composite faces in which 80 for searching and 20 for testing. For each of the 20 test faces, ten human objects are asked to select from the set of 80 composite faces the one that is the most similar to it and vote for that face. Then we can get 20 target faces for the 20 test faces based on the strategy of consensus voting [39] which chooses the one with the maximum number of votes. This result will be used as the standard to evaluate the performance of our encoding scheme. The set of 100 sampled composite faces is shown in Appendix A.

[5]It is due to the fact that our encoding scheme does not handle the size and position variation of facial features. Some composite images in the police database have these variations, so we ignore them. The problem of size and position variation will be discussed in next section.

The accuracy measure is based on the percentage of hits achieved where each hit is defined as ranking the desired target face as the top. Our method obtains an accuracy of 35%. If we treat a hit as ranking the desired target face within the top six, we can get an accuracy of 65%. The result is not good but acceptable. We think that it is mainly due to the subjective selection of target faces used in our evaluation. During the selection, we face a problem that for some test faces, it is difficult to find the most similar faces as the targets for them. Also, different people have different perceptual judgement. Therefore, such target faces may get just a few of votes but they are still selected based on the voting strategy. They may not be representative. Another factor may be the selection of face recognition technologies used to pre-compute the similarity between facial features. The current encoding scheme is mainly based on one face recognition technology, the LFA. It may be better to use more than one face recognition technology or a combination of them. Since the encoding scheme just makes use of simple calculations, the elapsed time in matching is negligible (it spends less than three seconds).

4.4 Shortcomings of the encoding scheme

During the process of facial recall, we may find that the selected feature templates may not exactly match with the corresponding facial features of the target person. Refinement is an inevasible process to make the selected feature templates look more similar to the corresponding facial feature of the target person. Refinement includes size variation and position variation. AICAMS-FIT allows such refinement: a feature template can be resized horizontally or vertically and it can be placed at arbitrary position within a specified boundary instead of the standard position. However, our

encoding scheme does not handle these kinds of variation. Therefore, we would like to investigate how such variations affect the matching. However, it is difficult to set up experiments to investigate the effect of position variation because facial features can be placed at arbitrary positions and it is unpredictable during the process of facial recall. Hence we have just performed experiments to investigate the effect of size variation because the size of facial feature templates varies in limited directions: either horizontal or vertical or both. Suggestions for solving these two problems are discussed in Chapter 6. In this section, we describe experiments to investigate the effect on matching if size variation occurs.

4.4.1 Size Variation

In our experiment, we prepare 20 artificial composite faces of human objects for the experiments by randomly selecting facial features of the six essential feature classes mentioned in Section 3.2.1 and combining them together. The 20 artificial faces are shown in Figure 4.7. For each feature class, we generate three sets of faces for it. The first set is generated by varying the width of the feature template of the testing class, the second is generated by varying the height of the feature template and the last is generated by varying both the width and the height of the feature template. Each set contains eight different extent of variations of faces, which are corresponding to the variation of -10%, -5%, -2%, -1%, +1%, +2%, +5% and +10%. Therefore, each test class has 24 test faces. Totally, there are 2900 faces in this experiment including the 20 original ones. Samples of the test faces are shown in Figure 4.8. These samples show a set of composite faces of the same object with various size of hair.

To investigate the effect of size variation, we use the change in recognition score as our measure. In our experiment, we use FaceIt Engine as our recognition tool again.

For each facial class, we match the original face of each object with its variations. Since the original face matches with itself must get a recognition score of 100[6], its variations should get a lower recognition score. Thus, the value of change in the recognition score indicates how each variation differ from the original one. This value can also be interpreted as the percentage change in the recognition score because the score ranges from 0 to 100. Actually, we are performing a sensitivity test on each single facial feature.

Our experimental results are shown in Tables 4.1-4.6: Table 4.1 shows the result when the size of hair is changing; Table 4.2 shows the result when the size of facial outline is changing; Table 4.3 shows the result when the size of eyes is changing; Table 4.4 shows the result when the size of eyebrows is changing; Table 4.5 shows the result when the size of nose is changing; Table 4.6 shows the result when the size of mouth is changing. The values in each table are the average of the results obtained from the matching of the 20 objects. Also, graphs are plotted to show the variation of recognition score against the variation of size. They are shown in Figure 4.9. Our observations from the results are described as follows:

1. For facial features such as facial outline and eyes, decreasing the size causes a greater change in the recognition score than increasing the size. For other features, both decrement and increment cause similar change in the recognition score.

2. For most of the facial features, the change of recognition score caused by varying both the horizontal size and vertical size is the greatest. Then followed by changing the horizontal size. Vertical size variation has the least effect on the

[6]The recognition score obtained from FaceIt Engine ranges from 0 to 100. It have been mentioned in Section 4.2.2.

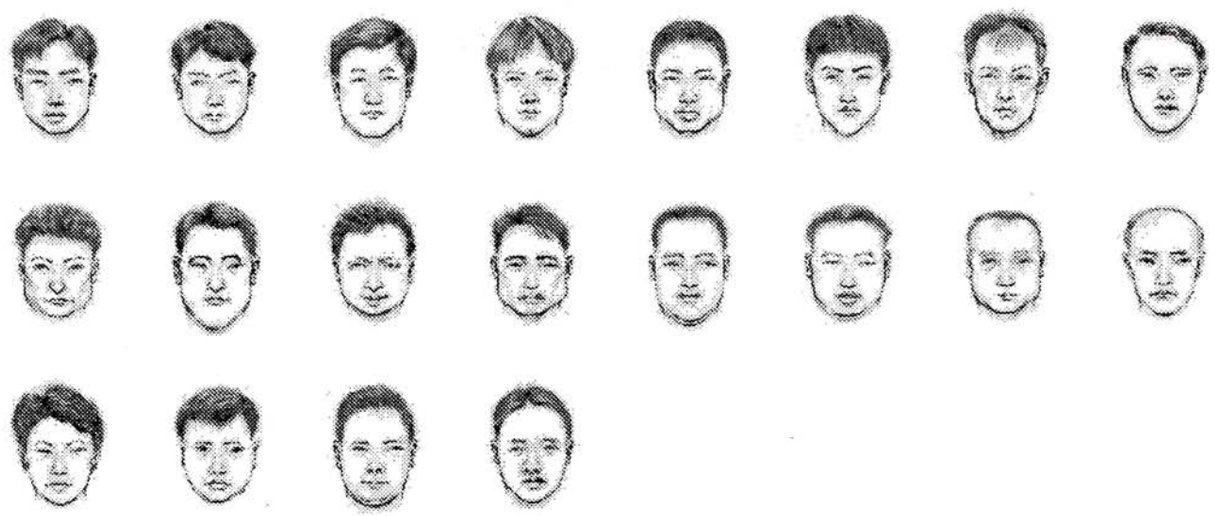


Figure 4.7: A database of artificial composite faces.

recognition score. However, for the mouth, vertical size variation has a greater effect on the recognition score than horizontal size variation.

3. The eyes are the most sensitive feature that affects the recognition performance because its variation changes the recognition score the most among other facial features. The sensitivity of other facial features is in the following descending order: facial outline, hair, eyebrows, nose and mouth.

The third observation of our result is consistent with the experimental analysis performed in [12]. That analysis is to investigate the discrimination power of each facial feature based on the recognition performance using that feature. Their result shows that among four single facial features (eyes, nose, mouth, whole face template), the eyes are the most discriminating template. This finding may help us to determine the value of ω for each facial feature in Equation 4.9.

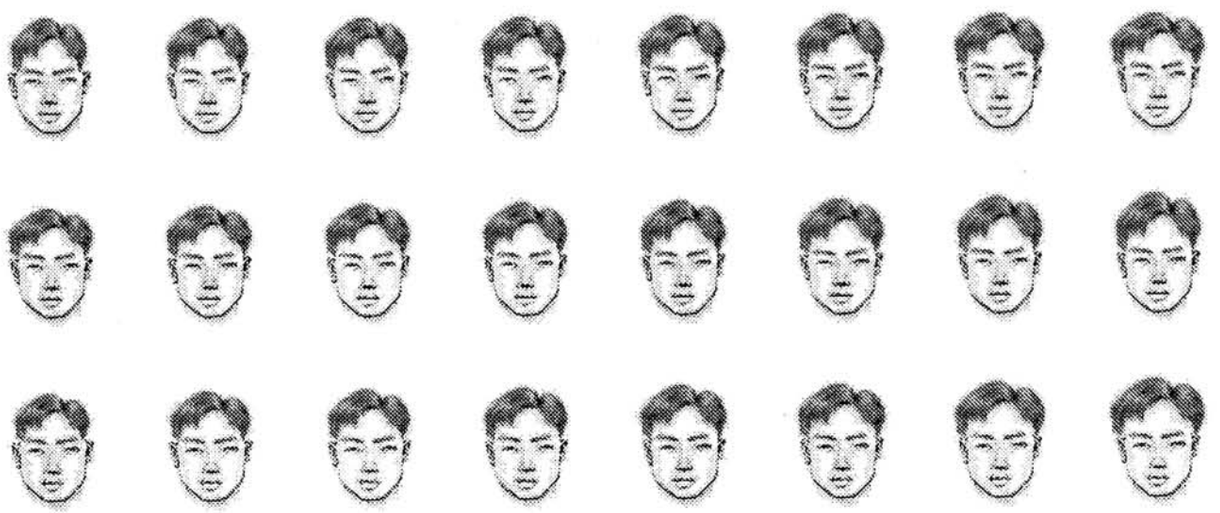


Figure 4.8: Sample of artificial composite faces with various size of hairs.

The first row shows composite faces with various horizontal sizes of hairs, the second with various vertical sizes and the last with various vertical and horizontal sizes.

Mode / Variation of Size	-1%	-5%	-2%	-1%	0%	+1%	+2%	+5%	+10%
Horizontal	4.35	3.45	2.55	1.75	0	1.85	2.2	2.9	3.7
Vertical	4.3	3.15	2.3	1.55	0	0.95	1.95	2.95	4
Both	4.4	3.5	2.35	1.35	0	1.3	2.3	3.1	4.35

Table 4.1: Effect on recognition score when the size of hair is changed.

Mode / Variation of Size	-1%	-5%	-2%	-1%	0%	+1%	+2%	+5%	+10%
Horizontal	4.9	3.25	2.2	1.3	0	1.15	1.6	2.85	3.75
Vertical	3.6	2.95	1.65	1.1	0	2.1	2.1	2.8	2.8
Both	5.6	3.7	2	1.4	0	2.15	2.35	2.5	2.95

Table 4.2: Effect on recognition score when the size of facial outline is changed.

Variation of Size \ Mode	-1%	-5%	-2%	-1%	0%	+1%	+2%	+5%	+10%
Horizontal	9.4	6.4	5.65	5.35	0	2.15	3.75	4.8	6
Vertical	5.3	3.75	2.8	1.45	0	1.4	2.35	3.75	4.7
Both	11.1	7.35	6.35	4.8	0	2.1	4.05	4.9	6.1

Table 4.3: Effect on recognition score when the size of eyes is changed.

Variation of Size \ Mode	-1%	-5%	-2%	-1%	0%	+1%	+2%	+5%	+10%
Horizontal	3.95	2.6	1.55	1.15	0	0.95	1.5	2.2	3.65
Vertical	3.2	2.7	1.8	0.9	0	1.25	1.5	1.65	2.2
Both	4.2	3.15	2.8	1.7	0	1.75	2.15	2.6	3.7

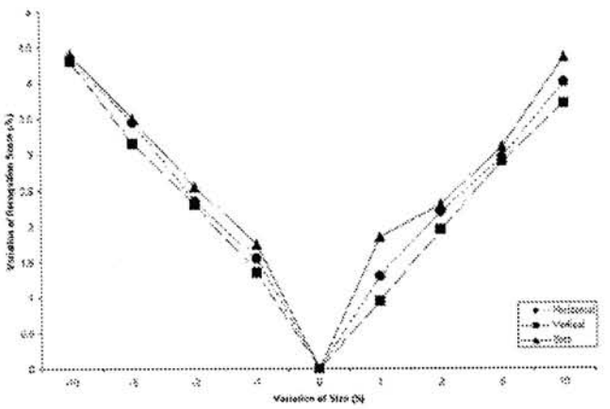
Table 4.4: Effect on recognition score when the size of eyebrows is changed.

Variation of Size \ Mode	-1%	-5%	-2%	-1%	0%	+1%	+2%	+5%	+10%
Horizontal	3.1	2.55	1.5	0.3	0	1.2	1.35	2	3.35
Vertical	3.65	2.15	1.6	0.85	0	0.8	1.4	2.15	2.7
Both	3.05	2.5	1.45	1.1	0	0.8	1.45	2	3.2

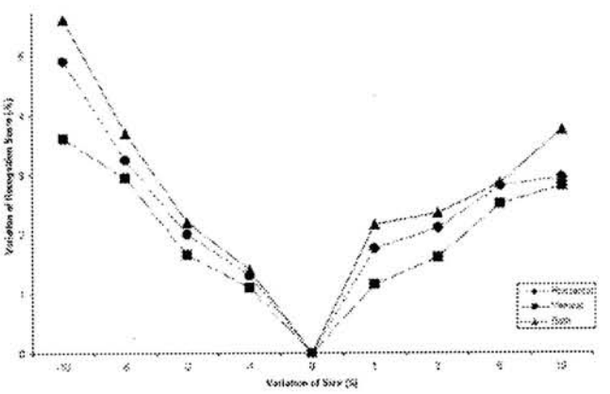
Table 4.5: Effect on recognition score when the size of nose is changed.

Variation of Size \ Mode	-1%	-5%	-2%	-1%	0%	+1%	+2%	+5%	+10%
Horizontal	2.35	1.65	1.15	0.55	0	0.45	1.1	1.45	2.3
Vertical	2.4	1.65	0.9	0.75	0	0.5	1.45	1.8	2.55
Both	2.7	1.65	1.3	0.95	0	0.55	1.3	2.4	2.6

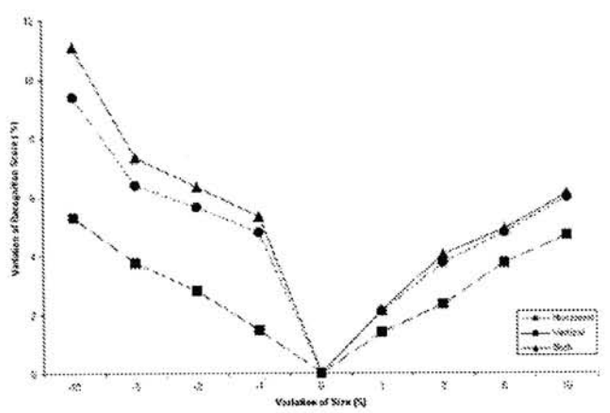
Table 4.6: Effect on recognition score when the size of mouth is changed.



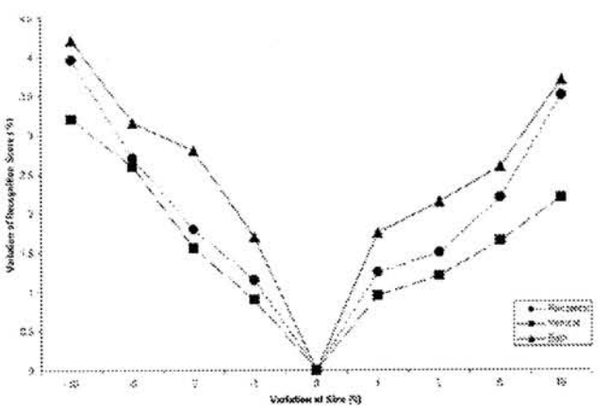
Hair



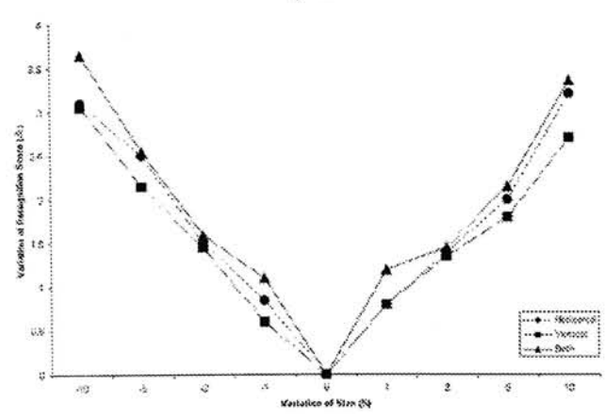
Facial Outline



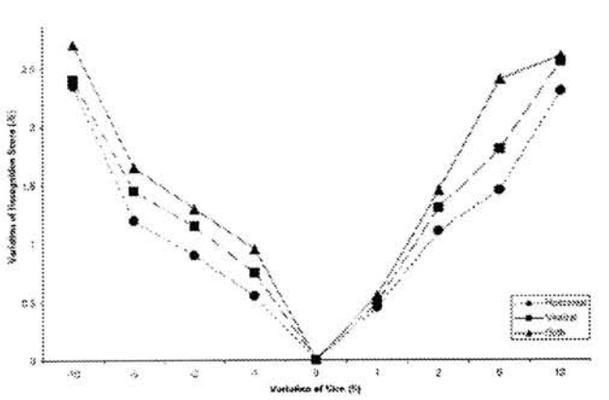
Eyes



Eyebrows



Nose



Mouth

Figure 4.9: Diagrams of size variation Vs recognition score.

4.5 Summary

In this chapter, a component-based encoding scheme is proposed for a recently used application of facial recall, AICAMS-FIT, to provide an efficient matching of facial composite sketches. The scheme makes use of the fact that each face is a composite of at least the six essential features. It builds a similarity matrix for each of the six essential features. Matching is performed by first comparing individual features and averaging the similarity values of each feature and selecting the maximum one. According to our evaluation method, the performance of the scheme is not good but acceptable. The current scheme has the limitations on size and position variations. After investing the effect of size variation on recognition, we find that eyes are the most sensitive facial feature among the others.

Chapter 5

Sketch-to-Photo/Photo-to-Sketch Matching

It is a common practice for law enforcement agencies to obtain facial photos of criminals when they are sentenced to be guilty. Usually, these photos are in frontal view and stored as grey-level images in a digital photo album. After facial recall, investigators may be interested to know if the photo album contains a photographic image of face that is similar to the reconstructed facial composite, which are made up of sketches of facial features as mentioned in Chapter 4. Thus, this involves a matching between facial photo and facial sketch provided that the facial composite is first converted into a holistic facial sketch[1]. In our work, we attempt to use two modern face recognition technologies, Principal Component Analysis (PCA) and Local Feature Analysis (LFA), in solving this kind of matching and do experiments to investigate their performances.

[1]It should be noting that the facial photo is a holistic facial image but the reconstructed face is a facial composite. Matching between a holistic face and a composite face is difficult. In AICAMS-FIT, this conversion is available.

5.1 Principal Component Analysis

Before going into the experiments, let us have a brief description about the two face recognition technologies, PCA and LFA. Since LFA has been mentioned in the previous chapter, we discuss PCA only in this section.

Principal Component Analysis (PCA), also known as Karhunen-Loeve expansion, has been applied in facial recognition by Turk and Pentland in [19]. In this approach, a facial image of size w by h is viewed as a two-dimensional $w \times h$ array of pixel values. Also, it may be considered as a vector of dimension $w \times h$ by a simple concatenation - the rows of the image are placed each beside one another [40]. In other words, the vector is belong to a point in an image space in which the dimension of all images is $w \times h$ pixels. Since facial images are similar in overall configuration, i.e. they all have two eyes, a mouth, a nose, etc. located at the same place, they are not randomly distributed in the image space and thus can be represented by a relatively low dimensional subspace. The aim of PCA is to build such a subspace which better describes the faces and it is called the face space. The basis vectors of this face space are called the principal components. Since they are face-like in appearance, they are also called "eigenfaces".

Let us consider a set of N facial images $\{ \Gamma_1, \Gamma_2, \ldots, \Gamma_N \}$ taking values in an v-dimensional image space, where $v = w \times h$. The average face of the set is defined by $\Psi = \frac{1}{N} \sum_{i=1}^{N} \Gamma_i$. Then each facial image differs from the average face by $\Phi_i = \Gamma_i - \Psi$. The covariance matrix W is defined as

$$W = AA^T \tag{5.1}$$

where $A = [\Phi_1 \Phi_2 \ldots \Phi_N]$. By PCA, W can be expressed as

$$W = \sum_{j}^{r} \lambda_j u_r u_r^T = U \Lambda U^T \tag{5.2}$$

where u_j is the jth eigenvector of W, λ_j is the eigenvalue of the associated jth eigenvector, and r is the rank of W. Besides, Λ is the $r \times r$ diagonal matrix of eigenvalues, U is the $v \times r$ matrix of eigenvectors in which eigenvectors are ordered according to their eigenvalues and $UU^T = I$ where I is the identity matrix. The eigenvector with the largest eigenvalue is referred to as the first eigenvector, which describes the direction of the largest variation from the average; the eigenvector with the second largest eigenvalue is referred to as the second eigenvector, which describes the direction of the second largest variation from the average, and so on [41].

Since the size of matrix W is v^2, determining the v eigenvectors is very time-consuming. A computationally efficient method is using the singular value decomposition of the matrix A:

$$A = U\Delta V^T \tag{5.3}$$

where U represents the matrix of eigenvectors of AA^T, V represents the matrix of eigenvectors of A^TA, and Δ is the diagonal matrix containing the singular values of A, which are defined as the non-negative square roots of the eigenvalues of AA^T which are the same of the eigenvalues of A^TA. Due to the fact that if the number of data points in the image space is less than the dimension of the space, i.e. $N < v$, there will only be $N - 1$ meaningful eigenvectors.[2] Thus, we only need to get $N - 1$ meaningful eigenvectors u_r of W. Generally, the number of images, N, is much less than the number of pixels in the images, v, i.e. $N \ll v$. In practice, a smaller $N' < N - 1$ is sufficient for recognition [19]. These N' eigenvectors are selected as those that have the largest eigenvalues which account for the most variance with the set of facial images and form a N'-dimensional face space. Having the N' eigenfaces (eigenvectors), a facial image Γ_{new} can be represented by transforming it into the face

[2]The remaining $v - N + 1$ eigenvectors will have associated eigenvalues of zero.

space as follows:

$$\omega_k = u_k^T(\Gamma_{new} - \Psi) \tag{5.4}$$

where $k = 1, \ldots, N'$. The weight ω_k is the projection of the face onto the kth eigenface which describes the contribution of kth eigenface in representing the input facial image. Hence, a facial image can be described by the weight vector $\Omega^T = [\omega_1 \omega_2 \ldots \omega'_N]$. Here, we can see the power of PCA: the dimension of face representation is reduced dramatically from v to N' where $N' < N \ll v$. A facial image can be reconstructed using its projected weight vector as follows:

$$\Phi_{rec} = \sum_{j}^{N'} \lambda_j \times \omega_j \times u_j \tag{5.5}$$

$$\Gamma_{rec} = \Phi_{rec} + \Psi \tag{5.6}$$

Facial Recognition is performed using two distance functions: the "distance in face space" (DIFS) and the "distance from face space" (DFFS). The former one is defined as the Euclidean distance between the weight vector of a novel face and the weight vector of a facial class [3]:

$$\epsilon_i = \| \Omega_{novel} - \Omega_i \|^2 \tag{5.7}$$

where Ω_i is a vector describing the ith face class which is the average of the weight vectors of all facial images of an individual. This distance is used to classify a novel face as either "known" or "unknown" by testing it with a threshold θ_ϵ. If $min\{\epsilon_i\} < \theta_\epsilon$, the novel face is said to belong to face class i. Otherwise, a new face class should be created.

The latter one is defined as the Euclidean distance between the novel image and the face space:

$$\epsilon^2 = \| \Phi_{novel} - \Phi_{rec} \|^2 \tag{5.8}$$

[3] A face class is an image class that contains all the facial images of an individual.

This is the squared distance between the mean-adjusted novel image and its reconstruction in the face space. Since many images can be projected into the face space, this distance is used to distinguish if the novel image is a facial image by comparing it with the threshold θ_ϵ, too. Similarly, if $\epsilon < \theta_\epsilon$, the image is classified as a facial image, otherwise, it is not. This distance metric is useful in the face detection [19] and the detection of facial features [23].

In [21] and [30], they use cosine distance instead of the Euclidean distance, so, equation 5.7 and 5.8 now become:

$$\epsilon_i = \frac{\Omega_{novel} \cdot \Omega_i}{\| \Omega_{novel} \| \| \Omega_i \|} \tag{5.9}$$

and

$$\epsilon = cos(\Phi_{rec}, \Phi_{novel}) = \frac{\Phi_{rec} \cdot \Phi_{novel}}{\| \Phi_{rec} \| \| \Phi_{novel} \|} \tag{5.10}$$

respectively.

In our work, we do not need to compute the DFFS because all the images used in the experiments are certainly human faces. In the forensic application of facial recall, we can also ensure the images used in its sequential matching process will be human faces only. Our implementation of PCA, therefore, only uses the DIFS as the similarity measure. Since two different equations, equation 5.7 and 5.9, are proposed for computing that distance, we do experiments for each equation and compare their performance in the matching.

5.2 Experimental Setup

In our experiments, we prepare two distinct sets of facial images in which one set contains facial photos and another contains facial sketches of the corresponding faces. The set of facial photos has only one facial image each of 100 individuals in which

59 of them are collected from the police and the rest of them are collected from AR Face Database[4]. Among the facial photos collected from the police, 21 of them are without background and the others with background. All the images collected from AR database are without background.

The set of facial sketches is provided by the police and is created by, first, selecting 21 facial photos (which are those without background and collected from the police.) from the photographic set and then asking a professional forensic artist to portray the corresponding sketches. So, the set of facial sketches has one facial image each of 21 individuals. The photo-sketch pairs of human faces are shown in Figure 5.2. Although, in each sketch, there is slight deviations from its corresponding photo, we can easily match them correctly with our perceptual judgement.

All the facial images collected are in frontal view and grey-level. They are then well aligned and scaled by applying affine transformation: For each face, the locations of both eyes are selected manually. They are then rotated so that the inter-eye line is horizontal and scaled down (with smoothing based on the scale factor) so that the inter-eye distance is 20 pixels. The fixed point of the faces is then set to be the middle of that line. Finally, the facial images are cropped through a 89×69 windows centered horizontally about the fixed point and starting 43 rows above it. The two sets of facial images are shown in Appendix B.1 and B.2.

[4]This face database is created by Aleix Martinez and Robert Benavente at the Computer Vision Center (CVC). It is publicly available and contains more than 4,000 color images. We have just selected only the frontal view of 41 human objects. The facial photos can be downloaded from URL: *http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html.*

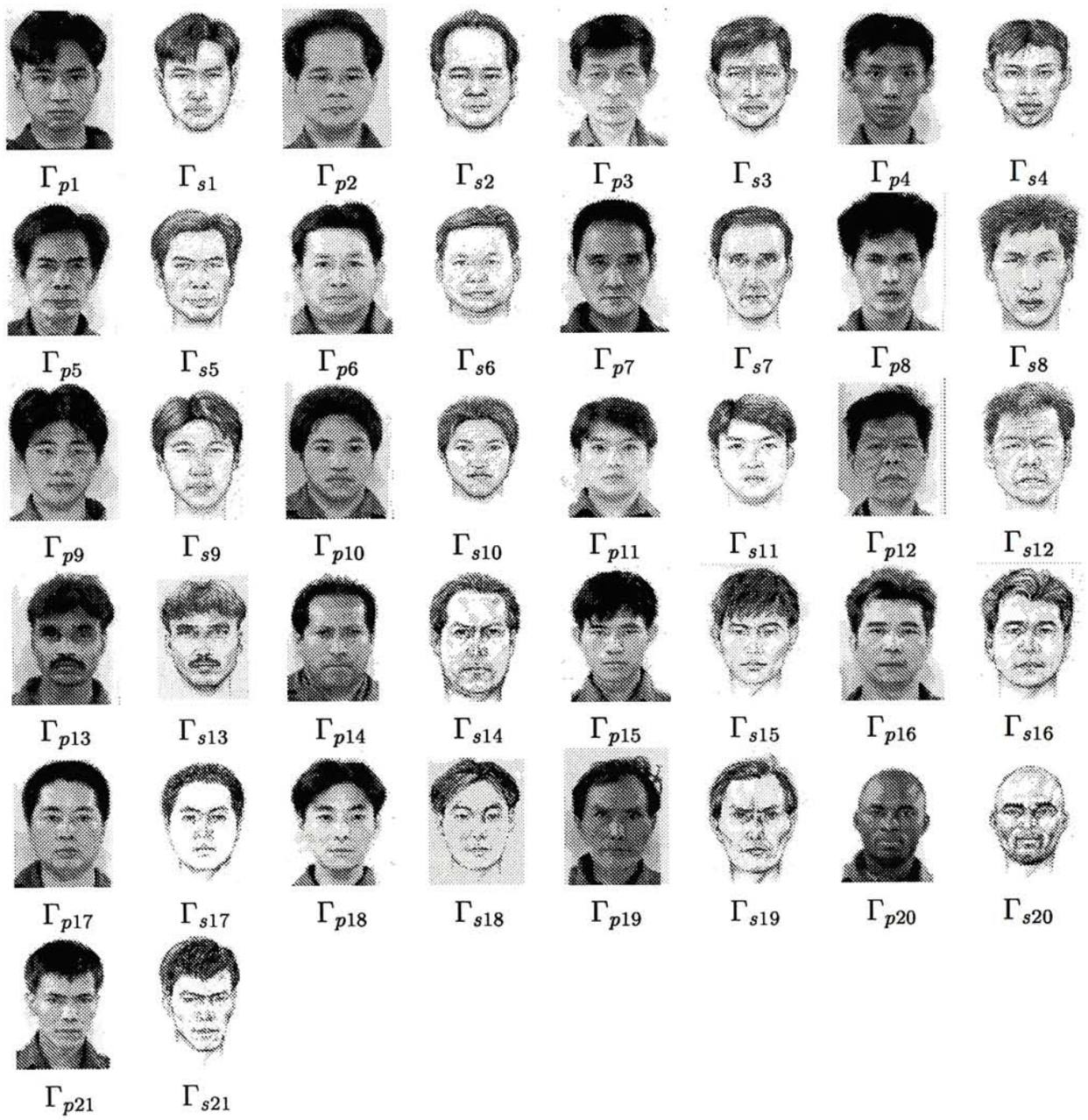


Figure 5.1: Photo-sketch pairs of facial images

5.3 Experimental Results

We have performed two kinds of facial matching: sketch-to-photo matching and photo-to-sketch matching. In the former matching, we use the set of photographic faces as the searching set while the set of sketches as the test set. Similarly, in the latter matching, the set of sketches, this time, is used as the searching set and the set of photographic images form the test set. Each time, a test image of face is compared with the searching set, if its corresponding image is selected as the top ranked one, it is considered as a hit. Recognition accuracy is measured based on the percentage of the hits achieved. In our experiments, we have a second measure of recognition accuracy which regards a hit as the target person is ranked within the top five. To implement PCA and LFA, we use MATLAB [5] and FaceIt Engine respectively.

5.3.1 Sketch-to-Photo Matching

For PCA, we perform experiments with the following three settings: let us use DB_{search} to denote the set of images for searching, DB_{test} to denote the set of images for testing and $T_{eigenfaces}$ to denote the number of eigenfaces used.

1. $DB_{test} = 21$ images of sketch, $DB_{search} = 21$ corresponding photos and $T_{eigenfaces} = 20$ (All).

   In this setting, we only use the photo-sketch pairs of facial images. Firstly, we generate 20 eigenfaces using the 21 facial photos to build an image space of facial photos. The eigenfaces generated are shown in Appendix C.1. Since the sample size is small, we can find that there is the in-sample effect as mentioned in [33]: It can be seen that most of the eigenmodes have very strongly expressed

[5] MATLAB is a registered trademark of The MathWorks.

individualities. For examples, the seventh eigenface u_{p7} looks like the fifth photographic face Γ_{p5} and the sixth eigenface u_{p6} looks like the second photographic face Γ_{p2}. We then project each facial sketch onto the space of photographic faces and perform the matching. The result is very poor, it achieves an accuracy of less than 30% for both Euclidean and cosine distance with the first accuracy measure while an accuracy of 31.10% and 47.62% for the Euclidean distance and cosine distance respectively with the second measure. We think that it is because facial sketch and facial photo are totally different in the intensity and PCA is sensitive to variations in intensity. However, when we reconstruct the facial sketches from the space of photographic faces, we discover that a few of the reconstructed faces look similar to their corresponding photographic faces although the reconstruction error is very high (about $2.7027e^{+3}$ on average). Figure 5.2 shows some of the examples, we can see that Γ_{s20}^{rec} can retain the bald head as same as its original image Γ_{s20} (or Γ_{p20}) in Figure 5.2.

2. $DB_{test} = 21$ sketches, $DB_{search} = 100$ photos, $T_{eigenfaces} = 53$

   In this setting, we use all the facial photos for searching to build the face space. Then we can just select the first 53 eigenfaces of largest eigenvalues as they explain about 95% variation of the image database [6]. This time, the in-sample effect becomes weaker. However, the performance becomes further poor, we can get only an accuracy of about 5% for both Euclidean and cosine distance. When we use the second measure, we get an accuracy of about 24% for both Euclidean and cosine distance. It is surprising that when we reconstruct the facial sketches,

[6] The number of eigenfaces required to explain ϵ variation is derived from the equation: $\frac{\sum_{i=1}^{T} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > \epsilon$ where N denotes the total number of eigenfaces. In our case, $N = 99$ and $\epsilon = 0.95$. Therefore, we get T = 53.

the reconstruction error is decreased by twice on average (about $1.3867e^{+3}$). More reconstructed faces look similar to their corresponding photographic faces and the similarity is higher. See Figure 5.3.

3. $DB_{test} = 21$ sketches, $DB_{search} = 100$ photos, $T_{eigenfaces} = 99$ (All)

   This setting is similar to the previous one, however, this time, we use all the eigenfaces generated from the photographic database. The performance is similar to the previous setting except that the Euclidean distance gets a little bit higher recognition accuracy of about 10% in the first measure. When we reconstruct the facial sketches, the reconstruction error further decreases (about $1.2210e^{+3}$). This time, most of the reconstructed faces look like their corresponding photographic faces. See Figure 5.4.

The performance of PCA with the above three settings is shown in Table 5.1.

For LFA, we perform experiments with two settings: 1. $DB_{test} = 21$ sketches and $DB_{search} = 21$ corresponding photos only. 2. $DB_{test} = 21$ sketches and $DB_{search} = 100$ photos. In both settings, the recognition performance is very good and we can obtain a perfect accuracy i.e. 100%. Although the average score obtained is not high (about 77.3), the engine ranks the corresponding photographic face as the top one when each facial sketch is presented. From Figure 5.6, we can easily find that LFA performs much better than PCA in this kind of matching.

Consider the case study in [3], they have done an experiment on sketch-to-photo matching with the setting of 13 facial sketches and 103 facial photos using a system called "PHANTOMAS". The evaluation method they used is different from ours. However, we can find that they obtain an accuracy of 61.54% and 84.62% with our first and second accuracy measures respectively. Compared to our result, we can see that LFA outperforms their system and their system is better than the PCA.

Γ^{rec}_{s11} Γ^{rec}_{s16} Γ^{rec}_{s18} Γ^{rec}_{s20} Γ^{rec}_{s21}

Figure 5.2: Sample of reconstructed facial sketches from photographic space ($N = 21$).

5.3.2 Photo-to-Sketch Matching

For PCA, we have only one setting for the experiment because we have limited number of facial sketches as the searching database. We generate 20 eigenfaces using the 21 facial sketches to build an image space of facial sketch. Since the sample size is small, there is also the in-sample effect as mentioned in the sketch-to-photo matching. We then project each photographic face onto the space of facial sketches and perform the matching. This time, the result is much better than that in Setting 1 of the sketch-to-photo matching. It achieves an accuracy of about 38% for both Euclidean and cosine distance. When we use the second measure, we get an accuracy of about 67%. However, when we reconstruct the faces, we find a similar phenomenon that the reconstructed facial photos look like their corresponding facial sketches. The reconstruction error, this time, is much higher (about $5.5073e^{+3}$) and the quality of reconstructed faces is poorer than that in the sketch-to-photo matching. See Figure 5.5.

For LFA, we obtain a recognition accuracy of 95.24% for the first measure. This time, the performance is a little bit worse than that in the sketch-to-photo matching. For the second measure, it achieves a perfect accuracy again. However, the LFA again outperforms the PCA. See Figure 5.7.

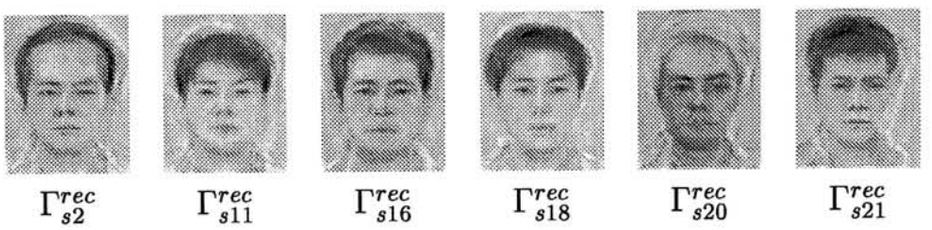


Figure 5.3: Reconstructed facial sketches from the image space of facial photos ($T_{eigenfaces} = 20$).

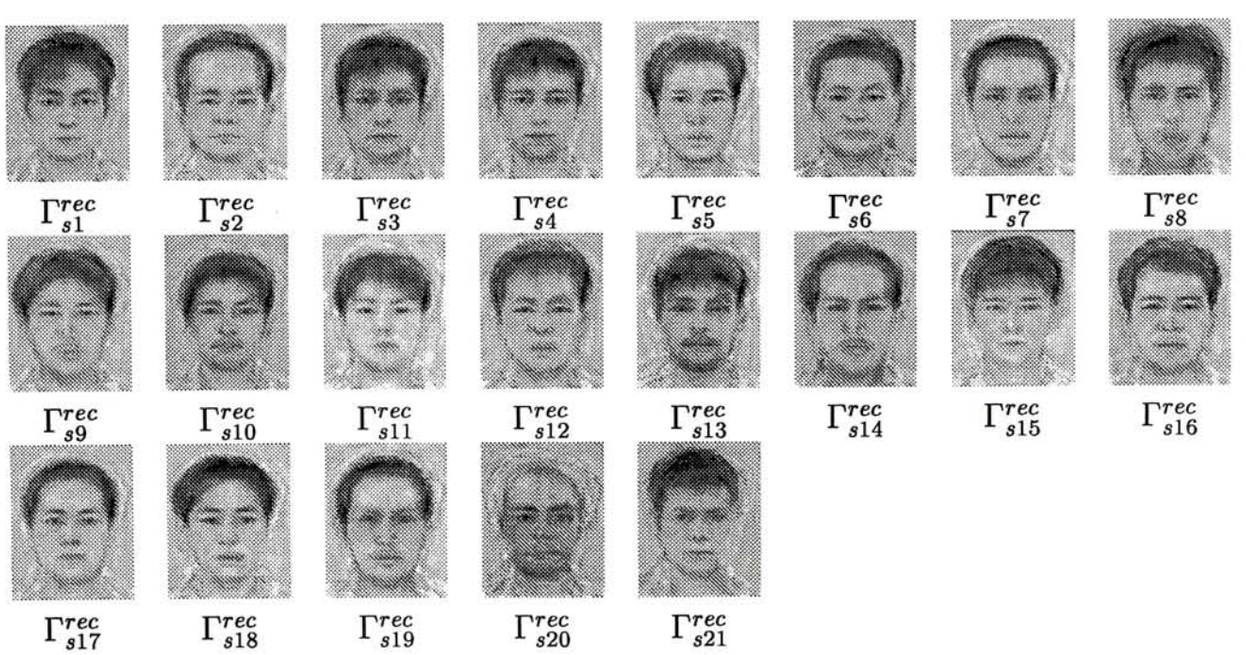


Figure 5.4: Reconstruction of facial sketches from the image space of facial photos ($T_{eigenfaces} = 99$).

We found that Γ_{s2}^{rec}, Γ_{s3}^{rec}, Γ_{s5}^{rec}, Γ_{s6}^{rec}, Γ_{s7}^{rec}, Γ_{s11}^{rec}, Γ_{s13}^{rec}, Γ_{s15}^{rec}, Γ_{s16}^{rec}, Γ_{s17}^{rec}, Γ_{s18}^{rec}, Γ_{s20}^{rec} and Γ_{s21}^{rec} look similar to their corresponding photographic faces shown in Figure 5.2.

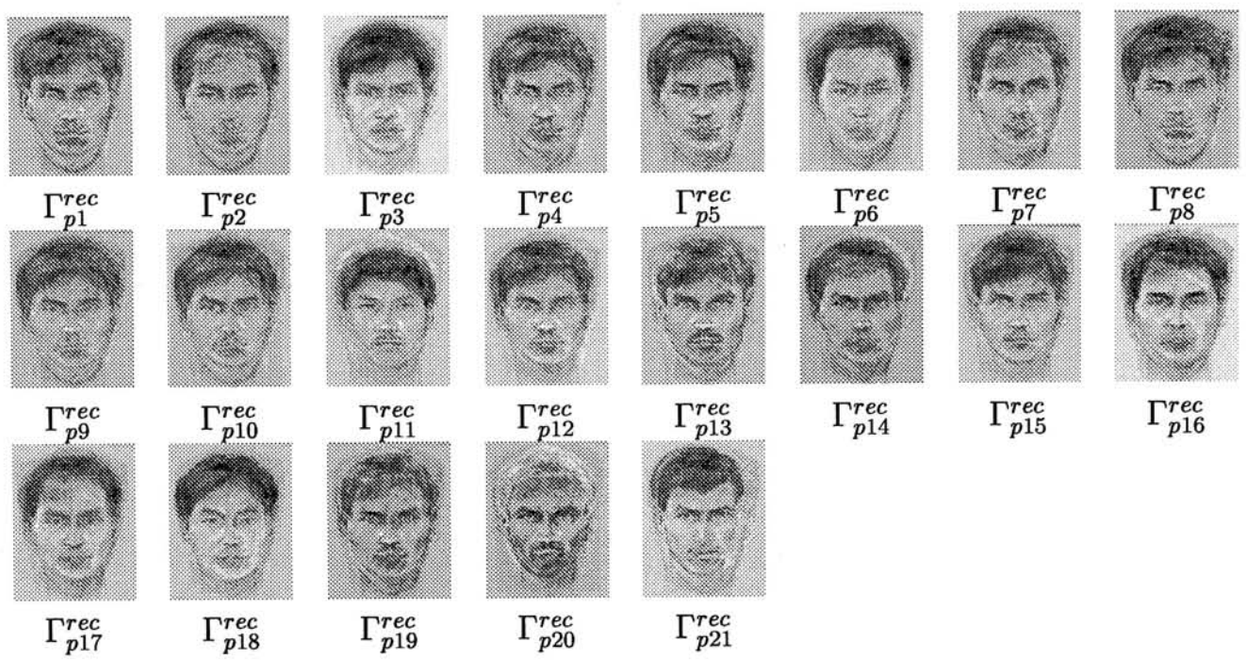


Figure 5.5: Reconstruction of facial photos from the image space of facial sketches ($T_{eigenfaces} = 20$).

We found that Γ_{p3}^{rec}, Γ_{p13}^{rec}, Γ_{s18}^{rec}, Γ_{s20}^{rec} and Γ_{s21}^{rec} look like their corresponding facial sketches shown in Figure 5.2.

	Rank = 1		Rank $\leq$ 5	
	Euclidean	Cosine	Euclidean	Cosine
Setting 1	23.81%	28.57%	38.10%	47.62%
Setting 2	4.76%	4.76%	23.81%	23.81%
Setting 3	9.52%	4.76%	23.81%	23.81%

Table 5.1: The performance of PCA with the three settings in the sketch-to-photo matching.

The first column (Rank = 1) of the table shows the result with our first accuracy measure: the target face should be ranked as the top one while the second column (Rank $\leq$ 5) shows the result with our second measure: the target face should be ranked within the top five.

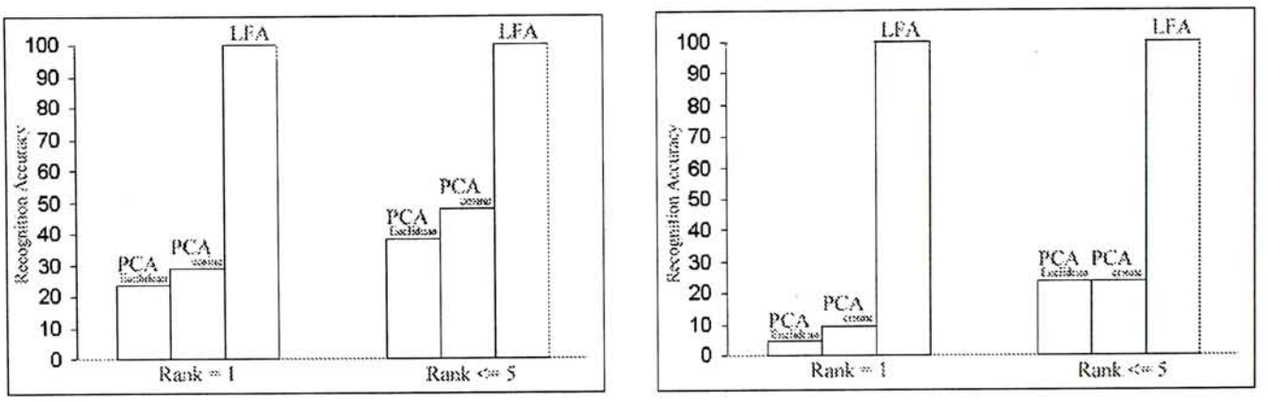


Figure 5.6: Comparison between PCA and LFA in the sketch-to-photo matching.

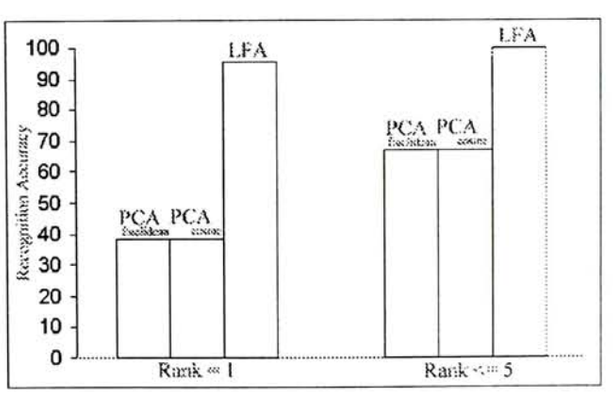


Figure 5.7: Comparison between PCA and LFA in the photo-to-sketch matching.

5.4 Summary

In this chapter, we have just attempted to test the performance of PCA and LFA in solving the problem of the sketch-to-photo/photo-to-sketch matching. According to our results, the performances between the two algorithms seem significantly different. LFA performs much better than PCA: LFA obtains a nearly perfect recognition accuracy while PCA obtains accuracy around 25% $\sim$ 35% for both matching. The performance gap is about 65% $\sim$ 75%. However, this is not the case when we consider the photo-to-photo matching. In [42], the FERET test reports that LFA outperforms the global PCA about 10% $\sim$ 20%. Why are the performance gaps between LFA and PCA in solving photo-to-photo matching and sketch-to-photo/photo-to-sketch matching so different? We have no clear answer to this question, but we think that it might be due to the following two factors: 1. The illumination of a sketch and a photo is significantly different and PCA is very sensitive to this factor, so its performance decreases dramatically. 2. LFA can produce receptive fields located at local features, such as nose, eye and the outlines of the face, as well as the edges. Since the dominant features of line drawings are the edges [3], so LFA can still successfully match a sketch with a photo or a photo with a sketch.

In our work, we also find that PCA performs better in the photo-to-sketch matching than the sketch-to-photo matching. Moreover, it is interesting to find that either the reconstructed facial sketches from the image space of facial photos or the reconstructed facial photos from the image space of facial sketches, they look similar to their corresponding facial photos or facial sketches respectively.

Chapter 6

Future Work

In our current work on the sketch-to-sketch matching, the proposed component-based encoding scheme has the limitations on size and position variation. Also, our work on the sketch-to-photo/photo-to-sketch matching is still in the experimental stage. More works can be done in these two areas. The followings are some suggested future works:

1. In the current implementation, our encoding scheme assumes that the feature templates selected will be placed at the default positions. Although these placements obey the topography of a general human face, different people will have slightly different locations of their facial features. For example, someone may have a small gap between his nose and his mouth, but someone may not. In other words, geometrical features such as distances between facial features should be taken into account. In AICAMS-FIT, the geometrical features can be obtained implicitly and semi-automatically with the fact that a composite face is constructed by combining different facial features stored in a facial component library. We can first mark the fiducial points of each facial feature template

manually and store this information into a database. The selection of fiducal points can be refered to [13]. Once a composite face is constructed, actual location of each fiducial point can be obtained through a simple translation. Distances between the fiducial points can then be measured automatically. Then each measured distance should be normalized using the inter-eye distance and a vector of the geometrical features of a composite face, v, can be obtained. Figure 6.1 shows an example of how to determine the locations of a nose in a composite face. Given the vectors of geometrical features, the similarity of the topography of two composite faces can be measured by:

$$S_{topography}(C_x, C_y) = \frac{v_x \cdot v_y}{\| v_x \| \| v_y \|} \tag{6.1}$$

where $S_{topography}(C_x, C_y)$ can further be transformed from $[1, -1]$ to $[0, 100]$. By adding the topography term into equation 4.10, the similarity measure between two composite faces becomes:

$$Sim(C_x, C_y) = \sum_{j=1}^{m} \omega_j \times Sim(P_{j,ki}, P_{j,kl}) - \lambda S_{topography}(C_x, C_y) \tag{6.2}$$

where $0 \leq \lambda \leq 1$ determines the relative importance of the topography term. Actually, the topography term also handles the problem of size variation because the size of each facial feature is already captured in each feature vector. Although the suggested method may require real-time comparison of geometrical features, the time should be negligible because fiducial points have already been detected and no feature detection is needed. We think this is one of the possible ways to handle the problem of size and position variations.

2. Our encoding scheme relies only on a single classifier, LFA, to measure the similarity between facial features. Other classifiers such as HMMs, Eigenface

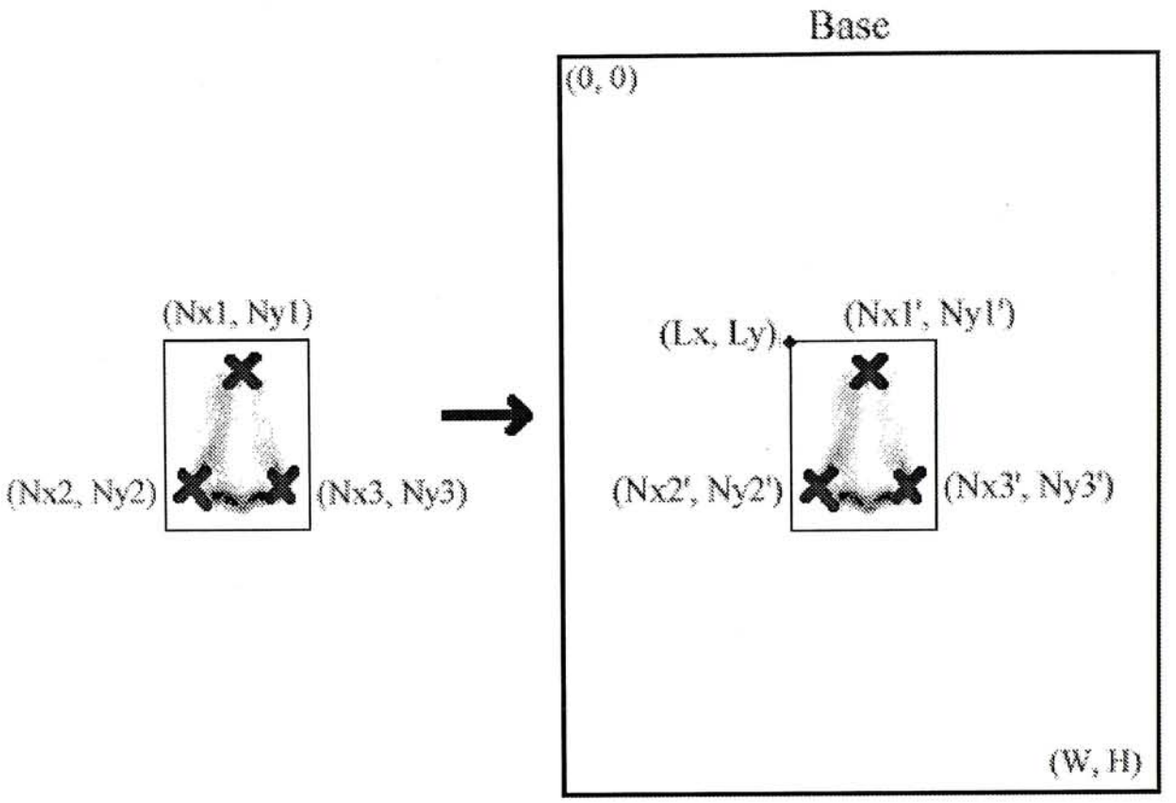


Figure 6.1: An example of determining the locations of facial features.

Given the feature template of a nose, we mark some fiducal points, $\{(N_{x1}, N_{y1}, (N_{x2}, N_{y2}), (N_{x3}, N_{y3})\}$, on it. When it is placed into a base template which forms the composite face, the locations of that fiducal points should now become $\{(N'_{x1}, N'_{y1}, (N'_{x2}, N'_{y2}), (N'_{x3}, N'_{y3})\}$ by $N'_{x1} = N_{x1} + L_x, N'_{y1} = N_{y1} + L_y$ where (L_x, L_y) is location of the upper left point of the nose template on the base template.

and Fisherfaces may be used. In [39], the authors show that the performance of a combined classifier is better than those of the single classifiers. The combined classifier is generated based on the output of several classifiers. It then yields a new ranking through a three-step fusion process: transformation of the score values, reduction of the combination set and combination and reordering. We can then transform the ranking into score and store them in the similarity matrix. Accuracy may be improved if we can use such combined classifier.

Chapter 7

Conclusions

Traditional facial matching technologies consider the photo-to-photo matching only. However, two new kinds of matching: sketch-to-sketch and sketch-to-photo/photo-to-sketch matching have been addressed in this thesis, relating to the problem of facial recall. For the sketch-to-sketch matching, we have proposed a component-based encoding scheme for a self-developed application of facial recall called AICAMS-FIT; For the sketch-to-photo/photo-to-sketch matching, we attempt to use two modern face recognition technologies, LFA and PCA, and compare their performance.

With AICAMS-FIT, a recalled face is represented as a composite of at least the six essential features obtained from a facial component library. Each facial feature in the library is actually portrayed by a professional forensic artist as a sketch of template. Our encoding scheme pre-computes the similarity between facial features of the same feature class using a robust facial recognition technology called the Local Feature Analysis. For each feature class, a similarity matrix is built. When a novel sketch-based composite face is presented and compared with others stored in the database, matching is performed by first comparing individual features and averaging

the similarity values of each feature and selecting the maximum one. According to our method of evaluation, the performance of the scheme is not good. However, the matching is efficient as it uses less than three seconds. Since the evaluation is based on human judgement, we think that the unsatisfied result may due mainly to the selection of target faces.

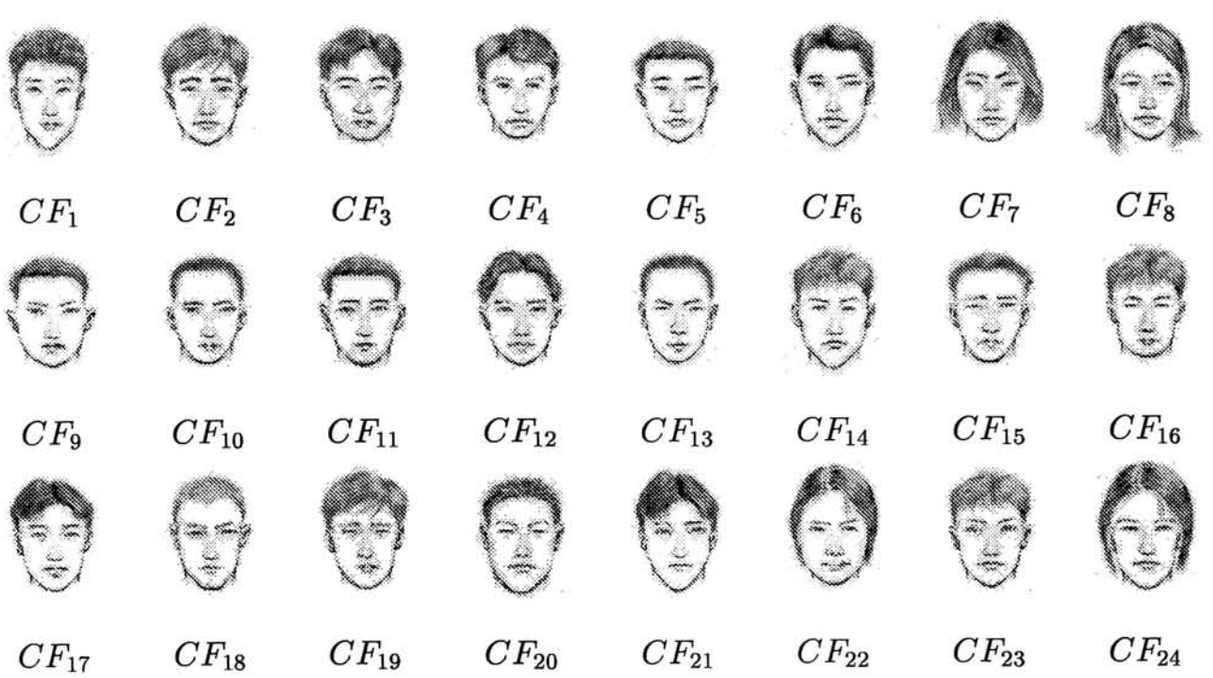
In the current scheme, we do not consider the case if the size or the position of a facial feature is changed. Experiments have been done to investigate the effect on recognition when size variation occurs. We discover that eyes' variation causes greater effect on recognition than other salient features. To overcome this problem, we suggest that geometrical features should be considered in addition to the current encoding scheme and a topography term is added into the current similarity measure. Also, we may use a combined classifier to improve the accuracy in measuring the similarity between facial features as well as the similarity between composite faces.

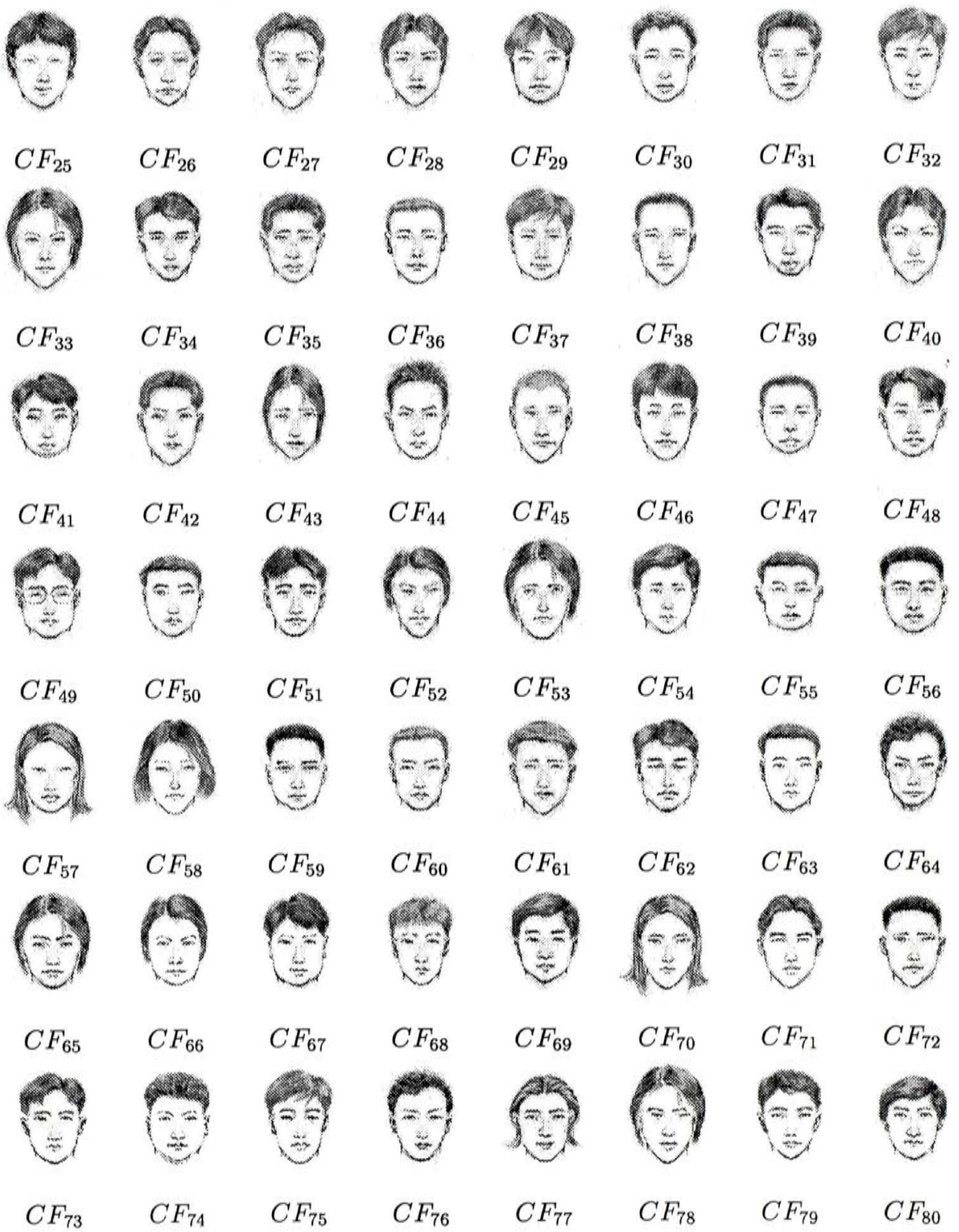
Moreover, we have done experiments to compare the performance of PCA and LFA in the sketch-to-photo/photo-to-sketch matching. Our results show that LFA performs much better than PCA and obtains a nearly perfect recognition accuracy with a database of 21 facial sketches and 100 facial photos. PCA performs a little bit better in the photo-to-sketch matching than the sketch-to-photo matching. An interesting finding has been obtained when we perform facial reconstruction: the reconstructed facial sketches from an image space of facial photos looks like their corresponding facial photos and the reconstructed facial photos from an image space of facial sketches looks like their corresponding facial sketches.

Appendix A

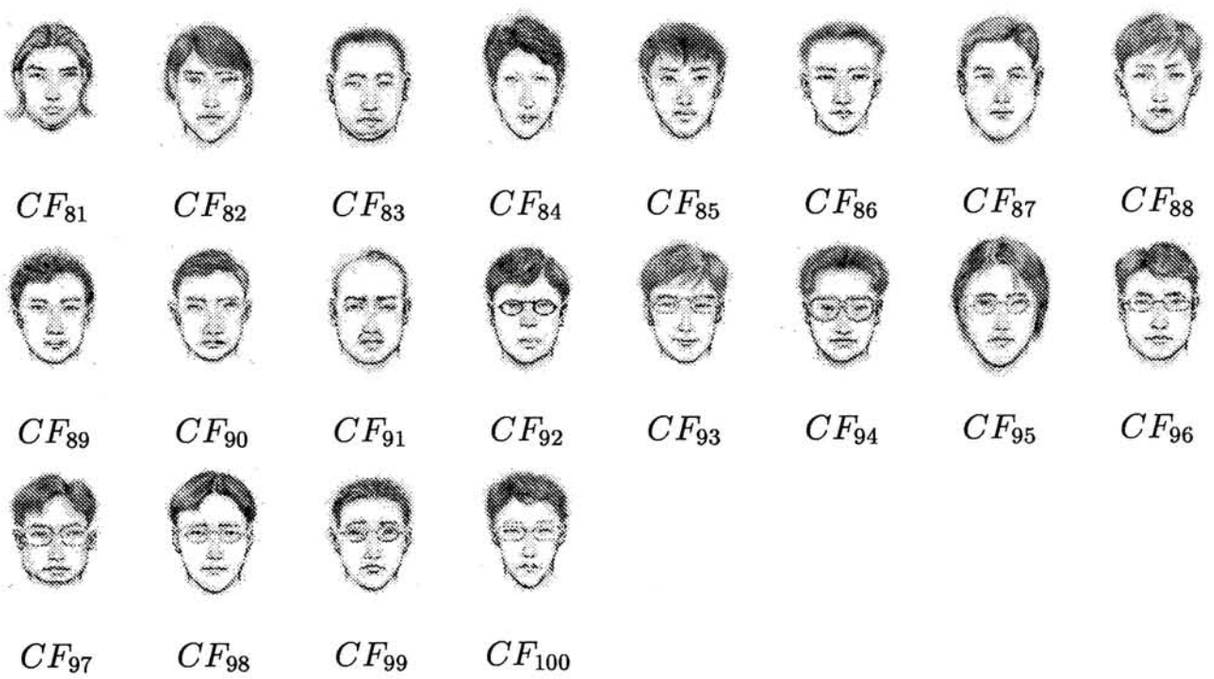
Image Library I

A.1 The Database for Searching




CF_{25} CF_{26} CF_{27} CF_{28} CF_{29} CF_{30} CF_{31} CF_{32}
CF_{33} CF_{34} CF_{35} CF_{36} CF_{37} CF_{38} CF_{39} CF_{40}
CF_{41} CF_{42} CF_{43} CF_{44} CF_{45} CF_{46} CF_{47} CF_{48}
CF_{49} CF_{50} CF_{51} CF_{52} CF_{53} CF_{54} CF_{55} CF_{56}
CF_{57} CF_{58} CF_{59} CF_{60} CF_{61} CF_{62} CF_{63} CF_{64}
CF_{65} CF_{66} CF_{67} CF_{68} CF_{69} CF_{70} CF_{71} CF_{72}
CF_{73} CF_{74} CF_{75} CF_{76} CF_{77} CF_{78} CF_{79} CF_{80}

A.2 The Database for Testing



CF_{81} CF_{82} CF_{83} CF_{84} CF_{85} CF_{86} CF_{87} CF_{88}

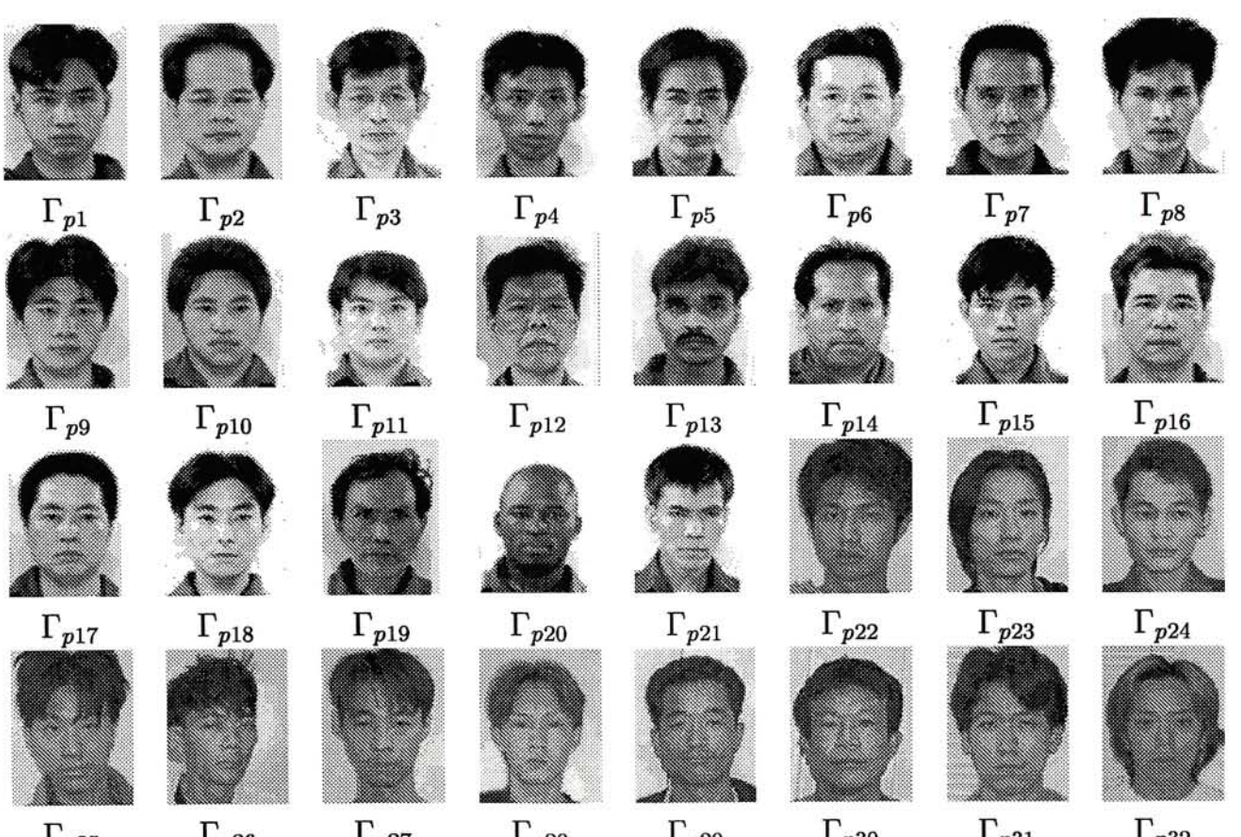
CF_{89} CF_{90} CF_{91} CF_{92} CF_{93} CF_{94} CF_{95} CF_{96}

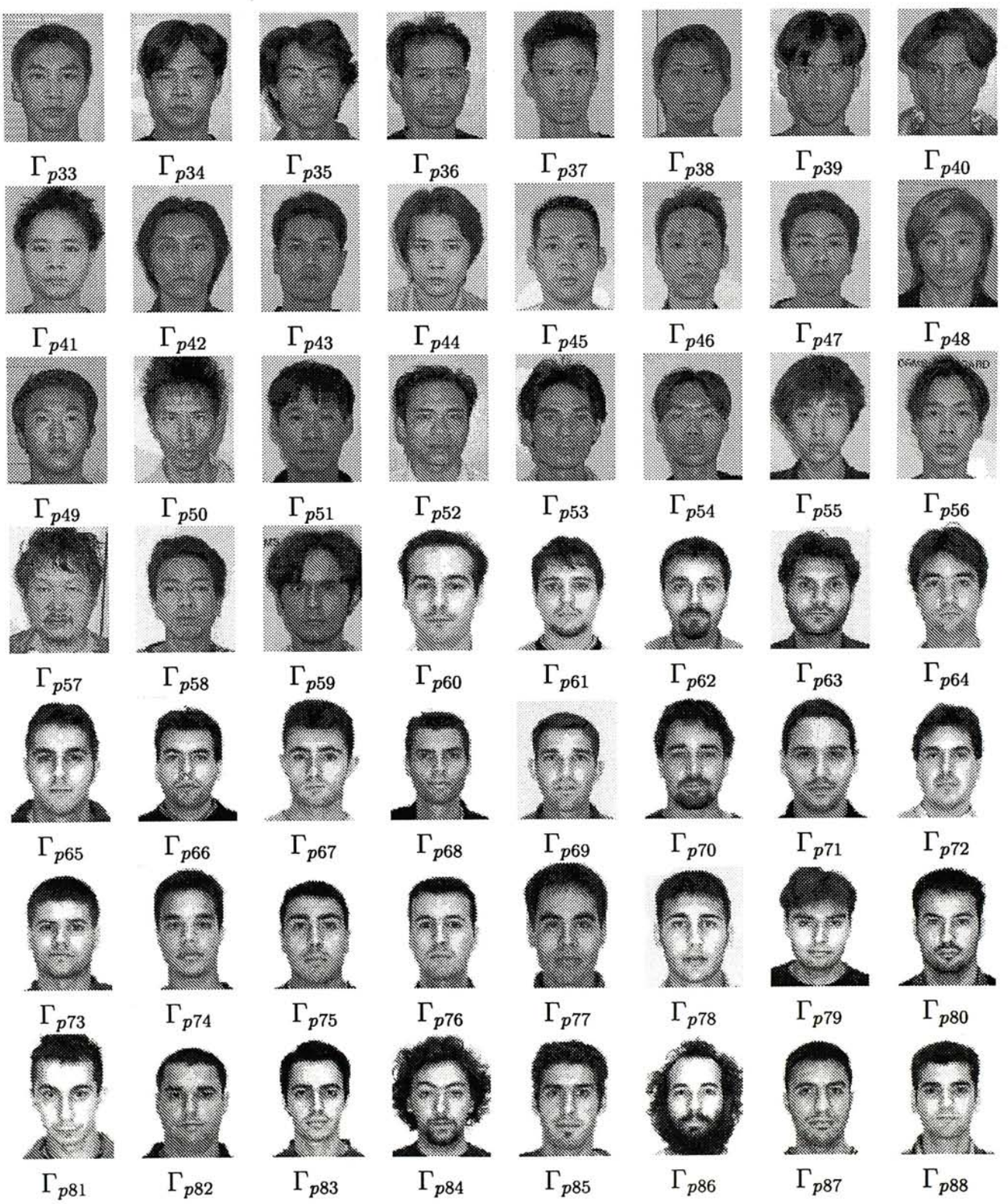
CF_{97} CF_{98} CF_{99} CF_{100}

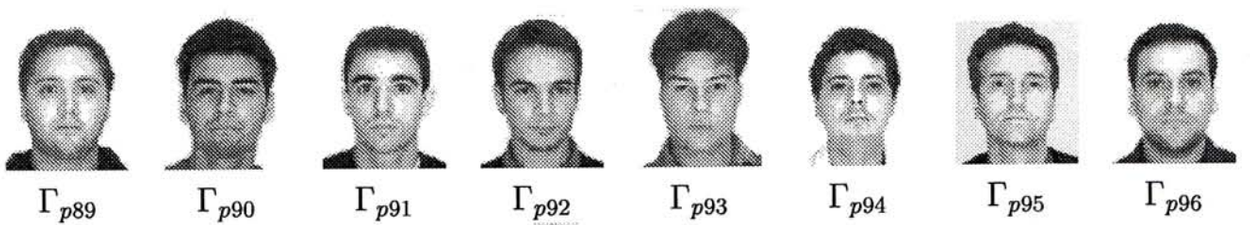
Appendix B

Image Library II

B.1 The Photographic Database




Γ_{p33} Γ_{p34} Γ_{p35} Γ_{p36} Γ_{p37} Γ_{p38} Γ_{p39} Γ_{p40}
Γ_{p41} Γ_{p42} Γ_{p43} Γ_{p44} Γ_{p45} Γ_{p46} Γ_{p47} Γ_{p48}
Γ_{p49} Γ_{p50} Γ_{p51} Γ_{p52} Γ_{p53} Γ_{p54} Γ_{p55} Γ_{p56}
Γ_{p57} Γ_{p58} Γ_{p59} Γ_{p60} Γ_{p61} Γ_{p62} Γ_{p63} Γ_{p64}
Γ_{p65} Γ_{p66} Γ_{p67} Γ_{p68} Γ_{p69} Γ_{p70} Γ_{p71} Γ_{p72}
Γ_{p73} Γ_{p74} Γ_{p75} Γ_{p76} Γ_{p77} Γ_{p78} Γ_{p79} Γ_{p80}
Γ_{p81} Γ_{p82} Γ_{p83} Γ_{p84} Γ_{p85} Γ_{p86} Γ_{p87} Γ_{p88}

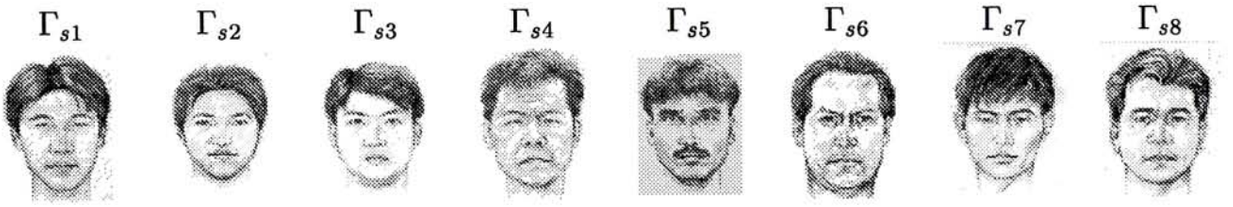


Γ_{p89} Γ_{p90} Γ_{p91} Γ_{p92} Γ_{p93} Γ_{p94} Γ_{p95} Γ_{p96}

Γ_{p97} Γ_{p98} Γ_{p99} Γ_{p100}

B.2 The Sketch Database

Γ_{s1} Γ_{s2} Γ_{s3} Γ_{s4} Γ_{s5} Γ_{s6} Γ_{s7} Γ_{s8}



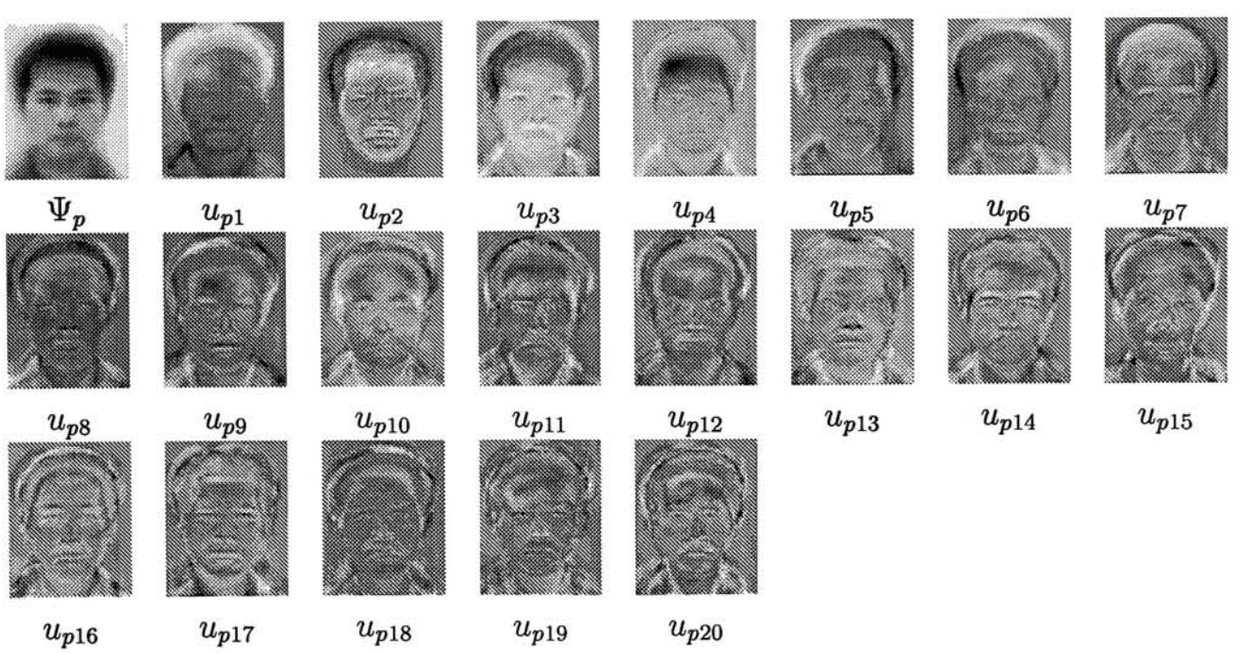
Γ_{s9} Γ_{s10} Γ_{s11} Γ_{s12} Γ_{s13} Γ_{s14} Γ_{s15} Γ_{s16}

Γ_{s17} Γ_{s18} Γ_{s19} Γ_{s20} Γ_{s21}

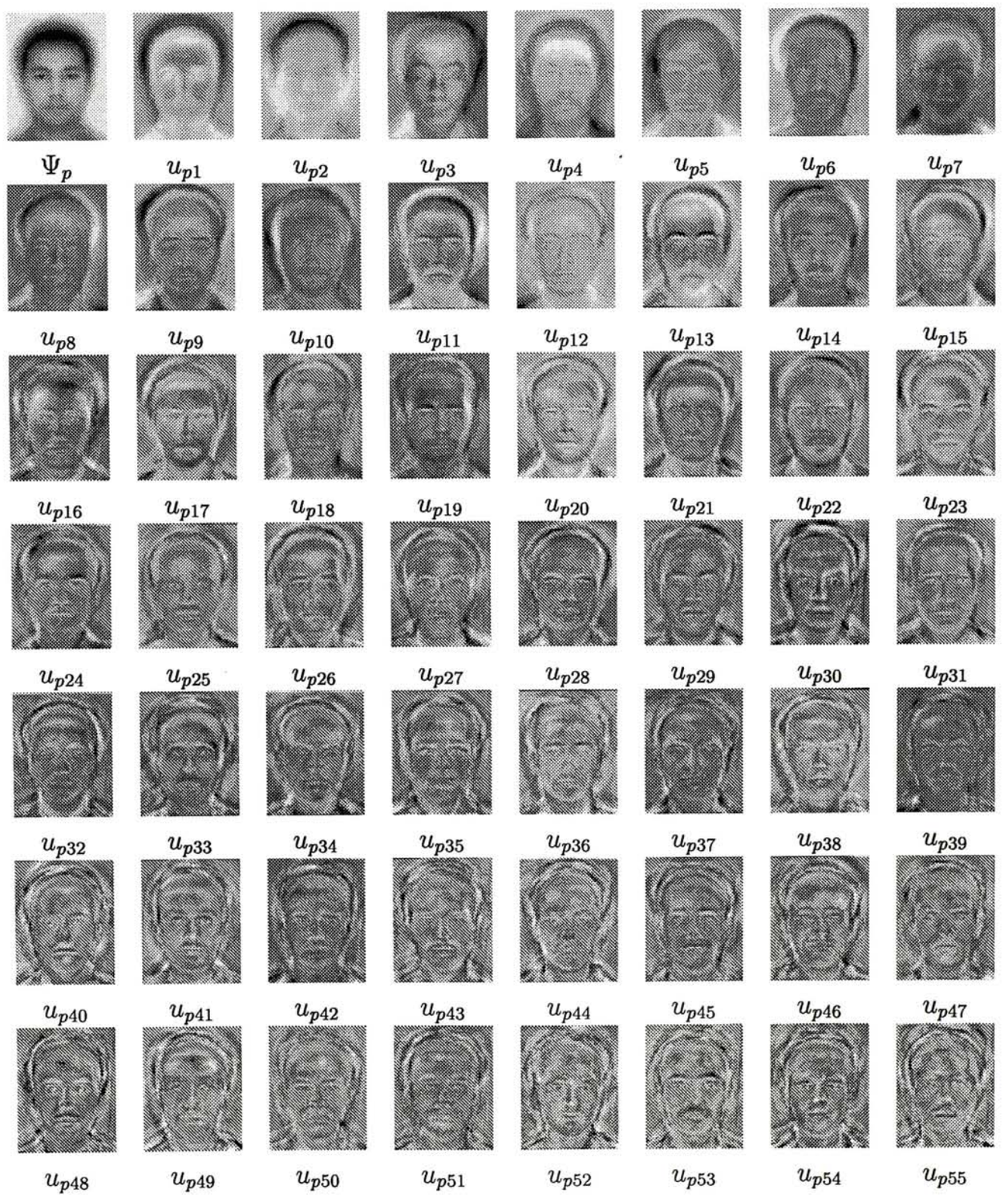
Appendix C

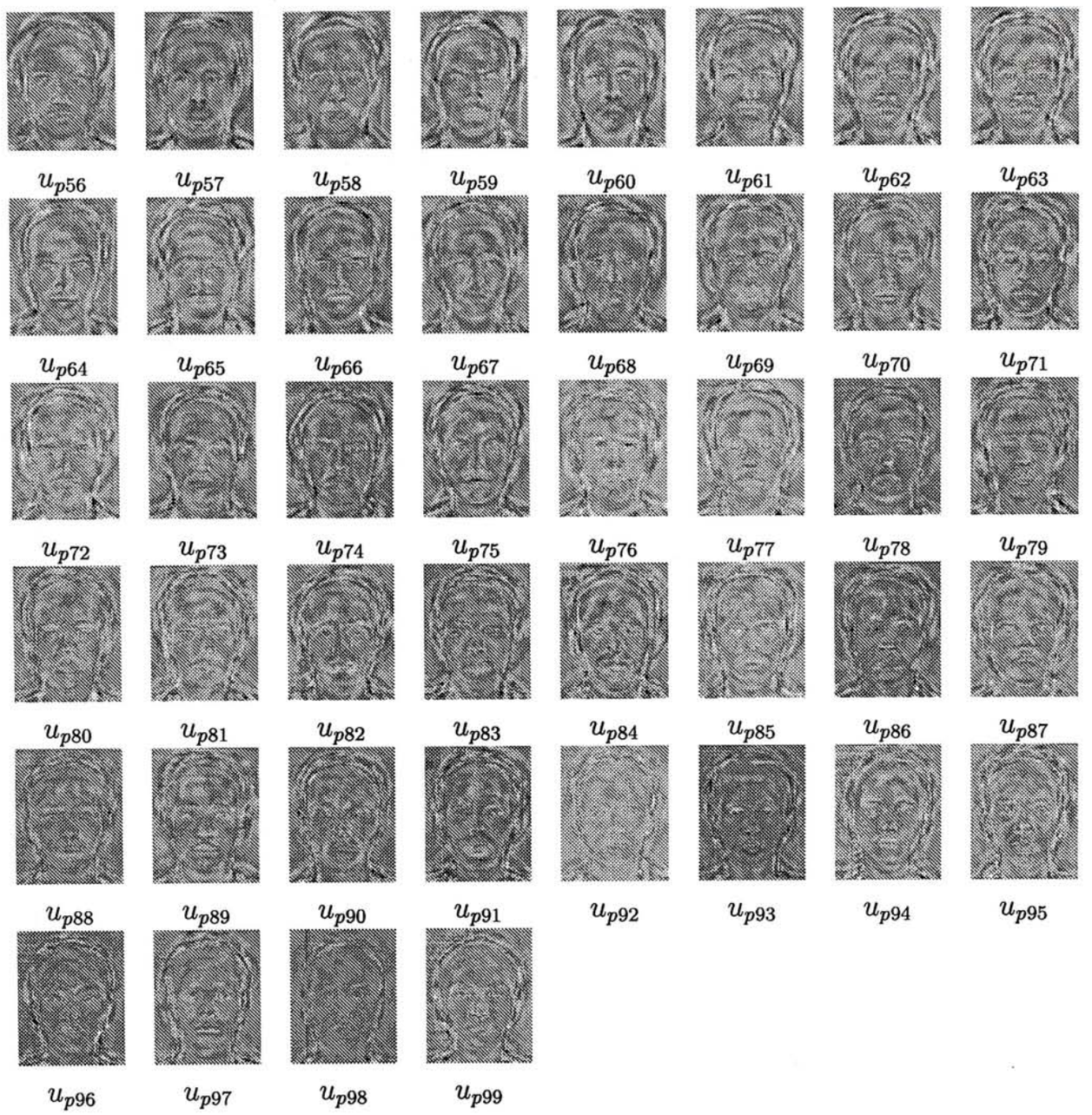
The Eigenfaces

C.1 Eigenfaces of Photographic Database ($N = 20$)

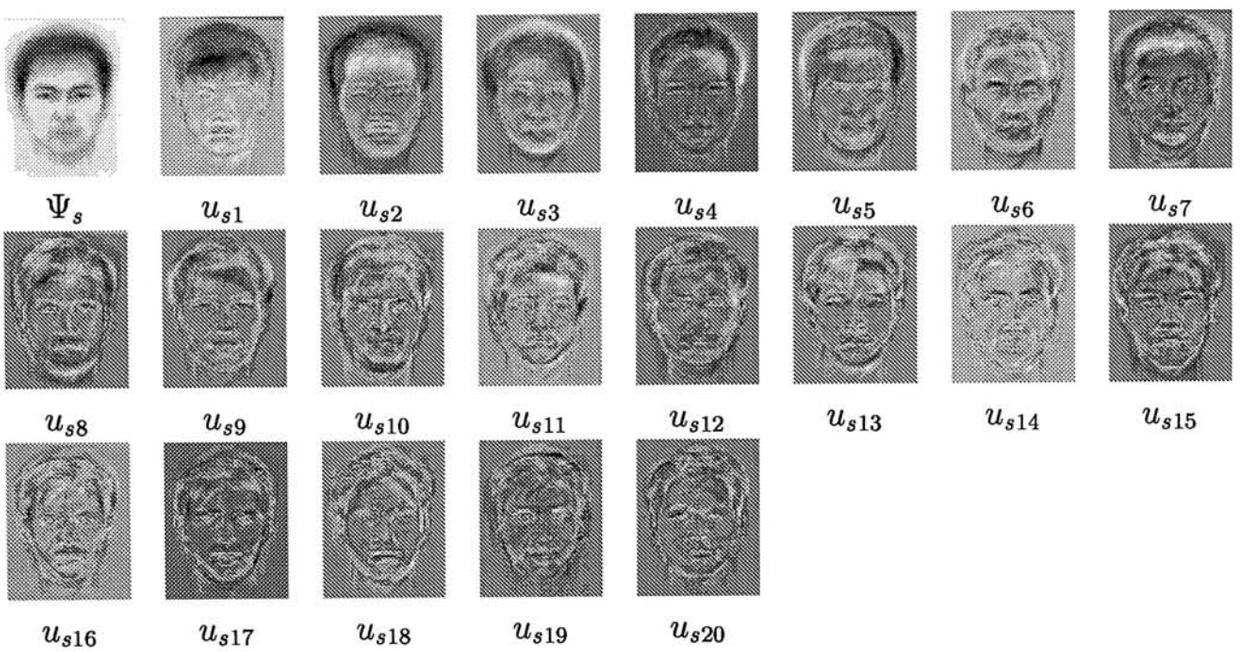


C.2 Eigenfaces of Photographic Database ($N = 100$)




u_{p56} u_{p57} u_{p58} u_{p59} u_{p60} u_{p61} u_{p62} u_{p63}
u_{p64} u_{p65} u_{p66} u_{p67} u_{p68} u_{p69} u_{p70} u_{p71}
u_{p72} u_{p73} u_{p74} u_{p75} u_{p76} u_{p77} u_{p78} u_{p79}
u_{p80} u_{p81} u_{p82} u_{p83} u_{p84} u_{p85} u_{p86} u_{p87}
u_{p88} u_{p89} u_{p90} u_{p91} u_{p92} u_{p93} u_{p94} u_{p95}
u_{p96} u_{p97} u_{p98} u_{p99}

C.3 The Eigenfaces of Sketch Database



Bibliography

[1] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of The IEEE*, vol. 83, pp. 705–740, May 1995.

[2] K. R. Laughery and M. S. Wogalter, "Forensic applications of facial memory research," in *Handbook of Research on Face Processing* (A. Young and H. Ellis, eds.), ch. 11, pp. 519–555, Elsevier Science Publishers B.V. (North-Holland), 1989.

[3] W. Konen, "Comparing facial line-drawings with gray-level images: A case study on phantomas," in *ICANN96*, (Bochum. Springer Heidelberg), 1996.

[4] H. Ellis, J. Shepherd, and G. Davies, "An investigation of the use of the photofit technique for recalling faces," *British Journal of Psychology*, vol. 66, no. 1, pp. 29–37, 1975.

[5] M. A. Mauldin and K. R. Laughery, "Composite production effects on subsequent facial recognition," *Journal of Applied Psychology*, vol. 66, no. 3, pp. 351–357, 1981.

[6] R. J. Phillips, "Recognition, recall and imagery of faces," in *Practical Aspects of Memory* (M. M. Gruneberg, P. E. Morris, and R. N. Sykes, eds.), pp. 270–277, London: Academic Press, 1978.

[7] G. M. Davies, "Forensic face recall: The role of visual and verbal information," in *Evaluating Witness Evidence* (S. M. Lloyd-Bostock and B. R. Clifford, eds.), pp. 103–124, John Wiley & Sons, 1983.

[8] W. W. Bledsoe, "The model method in facial recognition," tech. rep., Panoramic Research, 1964.

[9] Y. Kaya and K. Kobayashi, "A basic study on human face recognition," in *Frontiers of Pattern Recognition* (S. Watanabe, ed.), pp. 265–289, Academic Press, Inc., 1972.

[10] R. Brunelli and T. Poggio, "Face recognition through geometrical features," *Proceedings of ECCV '92*, pp. 792–800.

[11] T. Kanade, "Picture processing by computer complex and recognition of human faces," tech. rep., Kyoto University, Dept. of Information Science, 1973.

[12] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans on PAMI*, vol. 15, no. 10, pp. 1042–1052, 1993.

[13] I. J. Cox, J. Ghosn, and P. N. Yianilos, "Feature-based face recognition using mixture-distance," tech. rep., NEC Research Institute, 1995.

[14] J. Tojima and S. Sakamoto., *Private Communications.* 1994.

[15] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions On Computers*, vol. 42, pp. 300–310, Mar. 1993.

[16] L. Wiskott and C. von der Malsburg, "Face recognition by dynamic link matching," tech. rep., Institut für Neuroinformatik Ruhr Universität Bochum, Mar. 1996.

[17] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[18] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," in *7th Intl. Conf. on Computer Analysis of Images and Patterns, CAIP'97* (G. Sommer, K. Daniilidis, and J. Pauli, eds.), pp. 456–463, Sept. 1997.

[19] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[20] L. Sirovich and M. Kirby, "Low dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America*, no. 4, pp. 519–524, 1987.

[21] A. O'Toole, H. Abdi, A. D. Kenneth, and D. Valentin, "Low-dimensional representation of faces in higher dimensions of the face space," *Journal of the Optical Society of America. A, Optics and image science*, vol. 10, pp. 405–411, Mar. 1993.

[22] B. Moghaddam and A. Pentland, "Face recognition using view-based and modular eigenspaces," *Automatic Systems for the Identification and Inspection of Humans*, vol. 2277, July 1994.

[23] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," *IEEE Conference on Computer Vision and Pattern Recognition Conference*, pp. 84–91, 1994.

[24] C. Nastar and A. Pentland, "Matching and recognition using deformable intensity surfaces," Tech. Rep. 334, M.I.T Media Laboratory Perceptual Computing Section, 1995.

[25] C. Nastar, B. Moghaddam, and A. Pentland, "Generalized image matching: Statistical learning of physically-based deformations," Tech. Rep. 368, M.I.T Media Laboratory Perceptual Computing Section, 1996.

[26] B. Moghaddam, C. Nastar, and A. Pentland, "Bayesian face recognition using deformable intensity surfaces," Tech. Rep. 371, M.I.T Media Laboratory Perceptual Computing Section, 1996.

[27] B. Moghaddam, C. Nastar, and A. Pentland, "A bayesian similarity measure for direct image matching," Tech. Rep. 393, M.I.T Media Laboratory Perceptual Computing Section, 1996.

[28] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: Probabilistic matching for face recognition," Tech. Rep. 443, M.I.T Media Laboratory Perceptual Computing Section, 1998.

[29] M. S. Bartlett and T. J. Sejnowski, "Independent components of face images: A representation for face recognition," in *4th Annual Jount Symposium on Neural Computation*, May 1997.

[30] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski, "Independent component representations for face recognition," in *SPIE Symposium on Electronic Imaging:*

Science and Technology, pp. 529–539, Conference on Human Vision and Electronic Imaging III, 1998.

[31] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, July 1997.

[32] P. S. Penev and J. J. Atick, "Local feature analysis: a general statistical theory for object representation," *Network: Computation in Neural Systems*, vol. 7, pp. 477–500, Oct. 1996.

[33] P. S. Penev, *Local Feature Analysis: A Statistical Theory for Information Representation and Transmission.* PhD thesis, The Rockefeller University, May 1998.

[34] J. Goodman and E. F. Loftus, "Implications of facial memory research for investigative and administrative criminal procedures," in *Handbook of Research on Face Processing* (A. Young and H. Ellis, eds.), ch. 11.3, pp. 571–579, Elsevier Science Publishers B.V. (North-Holland), 1989.

[35] K. A. Deffenbacher, "Forensic facial memory: Time is of the essence," in *Handbook of Research on Face Processing* (A. Young and H. Ellis, eds.), ch. 11.2, pp. 563–569, Elsevier Science Publishers B.V. (North-Holland), 1989.

[36] L. Wiskott, "Phantom faces for face analysis," *Pattern Recognition*, vol. 30, no. 6, pp. 837–846, 1997.

[37] D. E. Pearson and J. A. Robinson, "Visual communication at very low data rates," *Proceedings of the IEEE*, vol. 73, pp. 795–811, 1985.

[38] C.-L. Huang and C.-W. Chen, "Human facial feature extraction for face interpretation and recognition," *IEEE*, 1992.

[39] B. Achermann and H. Bunke, "Combination of classifiers on the decision level for face recognition," Technischer Bericht IAM-96-002, Institut für Informatik und angewandte Mathematik, Universität Bern, Schweiz, Jan. 1996.

[40] S. RomdHani, "Face recognition using principal components analysis," Master's thesis, The University of Glasgow, 1996.

[41] D. Valentin, H. Abdi, B. Edelman, and A. J. O'Toole, "Principal component and neural network analyses of face images: What can be generalized in gender classification?," *Mathematical Psychology*, vol. 42, pp. 175–195, Dec. 1997.

[42] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.